All-or-None Transfer Based on Verbally Mediated Concepts

James G. Greeno

Indiana University, Bloomington, Indiana

AND

Joseph M. Scandura

University of Pennsylvania, Philadelphia, Pennsylvania

The problem of generalization and transfer is discussed relative to all-or-none learning, and a theory is proposed which combines features of the selection-of-strategies theory and the pattern-conditioning theory. A question is raised whether transfer of training involves recognition that must occur when a new item is first presented, or continued opportunities for transfer as long as the new item remains unlearned. This question is seen to involve different hypotheses about the initial and transition parameters of the Markov chain which describes the learning of items during a transfer task. Data are presented from an experiment in which positive transfer of training occurred based on generalization among members of a verbal concept-category. The data are consistent with the hypothesis that transfer occurred on the first presentation of a new item or not at all. The data are adequately fit by a theory expressed as a Markov chain in which learning is an all-or-none event. Statistical tests indicate that the effect of transfer can be described as a change in the initial vector of the chain, but that the learning parameters in the transition matrix varied little or not at all among conditions which differed widely in amounts of transfer. Finally, the amounts of transfer in 12 conditions are analyzed in relation to assumptions about the acquisition of verbal concepts and the recognition of new instances after a concept has been acquired.

This paper reports an analysis of a paired-associate memorizing task involving stimulus generalization between instances of verbal concepts. Verbal concept categories consist of words classified on the basis of an association test. The present study

1 This research was supported in part by Contract 908(16) between the Office of Naval Research and Indiana University, for which W. K. Estes was Principal Investigator, and in part by grants GB-319 and GB2791 from the National Science Foundation. The second author held a National Science Foundation College Teacher Fellowship while working on this project. We are grateful to Mrs. Alice Scandura who assisted in running the subjects, and to Mr. Larry Fritschel who assisted in the analysis of data. Mr. Fritschel's participation in the project was made possible by the Indiana High School Student Science Institute.
used results collected by Underwood and Richardson (1956) where associations were restricted to sense-impressions such as “red” or “large.” The instances of a given concept are words to which a substantial number of subjects gave the concept response as their association.

In a common experimental setup several verbal concepts are represented in a list. The list includes some number of instances of each concept. The subject’s task may be to learn the appropriate concept label for each item in the list, or the items may be paired with other verbal responses such as nonsense syllables for paired-associate learning. In the latter case, all the instances of a given concept would be paired with the same response. In either case, the task consists of learning a single response for all the instances of a concept while learning other responses to the instances of other concepts in the list.

Mastery of a verbal concept task is facilitated by generalization of the correct response from learned instances to other instances of the same concept. Consider a case in which the subject knows none of the stimulus-response associations of the experiment when training begins. The subject’s first accomplishment might be to learn the association between one stimulus in the concept category and its correct response. After one association is learned, the response can generalize from the first-learned member of the concept category to other instances and thereby facilitate learning of the later items. Presumably the generalization effects would increase or become more probable as the subject acquired additional associations involving concept instances.

We are accustomed to thinking about stimulus generalization in terms of amounts associative strength transferred from one item to another. However, a theory which postulates graded amounts of generalized strength apparently presupposes a theory of learning involving a gradual buildup of associative strength. If we are to take account of recent evidence supporting an all-or-none theory of learning (Bower, 1961; Estes, 1960, 1964; Rock, 1957) it is necessary to develop a compatible theory of generalization and transfer.

**General Considerations Regarding Transfer**

Theories assuming all-or-none learning have been extended in two ways to apply to generalization. The two extant theories of generalization supply complementary frameworks for considering the generalization problem.

First, consider the pattern-conditioning theory of learning and the pattern-components theory of generalization. The learning theory has been developed mainly by Estes (1959, 1960, 1964) and Bower (1961, 1962). The application to generalization is due mainly to Friedman (1962) and Friedman and Gelfand (1964) with an important theoretical development given by Atkinson and Estes (1963, pp. 243-252). The theory assumes that a stimulus pattern becomes connected to a response in an all-or-none
fashion. Then if a new stimulus is presented which contains one or more components in common with the first stimulus, transfer of the response from the first stimulus will occur. To illustrate: suppose that the subject learns a stimulus-response pair that can be represented abstractly as \( (abc) \rightarrow R \). According to the theory, this learning achievement consists of connecting the response \( R \) to the entire stimulus pattern \( (abc) \). Now suppose that another pair is presented; represent the second pair as \( (ad) \rightarrow R \). The pattern-components theory asserts that transfer will occur if the subject recognizes the presence of component \( (a) \) in both stimulus patterns.

An alternative theory which also assumes all-or-none learning is the selection-of-strategies theory due to Restle (1961, 1962). Important extensions and application to concept identification were given by Bower and Trabasso (1964) and Trabasso and Bower (1964). A further development and application to paired-associate learning was worked out by Greeno (1966). Stimulus generalization effects have been examined with this theory by Polson, Restle and Polson (1965), Restle (1964a, 1964b) and Trabasso (1963). The theory assumes that learning is selective—that is, the subject’s mastery of an item may involve an association between the response and a part of the stimulus. Then transfer to a new item occurs if the strategy selected by the subject for the first item also is an adequate strategy for remembering the response to the second item. To illustrate again: let the initial and transfer stimulus-response pairs be represented \( (abc) \rightarrow R \) and \( (ad) \rightarrow R \), as before. The subject is assumed to learn the first item by connecting response \( R \) with some selected aspect or combination of aspects of the pattern \( (abc) \). One such strategy would be to connect \( R \) to \( (a) \), but there are other possibilities, such as connecting \( R \) to \( (b) \) or connecting \( R \) to the combination \( (ac) \). Transfer to \( (ad) \rightarrow R \) will occur if the strategy selected earlier will support memory of the new item as well. In the present example, the selection of \( (a) \rightarrow R \) as the strategy would lead to transfer, but other strategies would not.

These two theories are genuinely complementary, and can be combined to yield a single, general theory of learning and transfer. The selection-of-strategies theory is a generalization of the pattern-theory of learning. According to the pattern theory, the entire stimulus pattern becomes conditioned to the correct response at some time. After that time, the item is learned. According to selection-of-strategies theory, either the entire stimulus pattern or some selected aspect of the pattern becomes associated with the correct response at some time, and after that time the item is learned. The selection-of-strategies theory thus is more general with respect to learning, since it allows for the greater number of possibilities.

On the other hand, the pattern-components theory is a generalization of the selection-of-strategies theory of transfer. According to selection-of-strategies theory, transfer will occur if the strategy selected for the first item also will support mastery of the second item. According to the pattern theory, transfer will occur if the second item has a component in common with the subject’s memory of the first item. Thus, the pattern theory allows for transfer in all the cases where the selection-of-strategies
theory does, and some additional cases as well. Specifically, the pattern-components theory suggests that transfer will occur of the initial item is learned with a conjunctive strategy involving an aspect of the second item. If \((abc)\rightarrow R\) were learned with a strategy \((ab)\rightarrow R\), transfer to \((ad)\rightarrow R\) would occur according to the pattern-components theory, but not according to selection-of-strategies theory.

It seems reasonable to desire the more general theory of both learning and transfer, and the generalization is conceptually straightforward. Assume that learning occurs whenever the subject succeeds in associating any stable aspect or combination of aspects of the stimulus with the correct response. The selected set of stimulus aspects may be called the subject’s code of the initial stimulus. Now suppose that a second item is presented such that the initial response is to be learned for a new stimulus. Assume that transfer will occur if the new stimulus has one more aspects in common with the subject’s code of the first stimulus. To use the formal illustration once more, the subject may code \((abc)\rightarrow R\) in a number of ways for initial learning. The code may be \((a)\rightarrow R\), \((ac)\rightarrow R\), \((c)\rightarrow R\), \((bc)\rightarrow R\), etc. Now when \((ad)\rightarrow R\) is presented for learning, transfer will occur if the first item was coded in any way that includes aspect \((a)\), either by itself or in combination with other aspects.

The general theory, outlined above, provides the basis for the present analysis. However, one additional extension is required because of the nature of the materials used. When we apply these ideas to the learning of verbal concepts, we include as aspects of the stimulus certain associative responses which the subject can supply. For example, if a stimulus is the word DIME, the “pattern” potentially includes such “components” as shiny, round, and flat, even though none of these would be used to describe the physical configuration of letters which the subject sees in the experiment. When materials like this are used, the presence or absence of a component in a pattern is a probabilistic matter; the association “round” may occur to some subjects when they see DIME but it probably will not occur to all subjects. This matter can be handled by including a recognition process in the theory. Given the initial and transfer items \((abc)\rightarrow R\) and \((ad)\rightarrow R\), transfer depends on two factors: (1) the subject must code the first item in a way that includes the component \((a)\), and (2) the subject must then recognize the presence of component \((a)\) as a part of the transfer item.

The above account relates to the conditions in which transfer will occur. It leaves unanswered certain questions about how the learning of new items is affected by transfer when it occurs. There are two main possibilities, which we can call first-trial recognition and continuing opportunities. According to the first-trial recognition hypothesis, the critical trial for transfer is the first presentation of an unlearned item after an initial similar item has been learned. If the subject recognizes a common aspect on that trial, transfer will occur. But if a common aspect is not recognized on the first trial, then the new item must be learned in the way that it would be if the first item had never been learned. The continuing opportunities hypothesis would permit the subject to transfer the response from a learned item to an unlearned similar item
as long as the second item remained unlearned. As we will see below, when the learning process is represented as a Markov chain, the two hypotheses lead to different expectations about the effect of transfer on the parameters of the chain.

**Formal Statement of Assumptions**

Assumptions about transfer can not be stated formally or tested at all except in conjunction with assumptions about learning. The learning assumptions to be applied here have been presented and analyzed in detail in other papers (Atkinson and Crothers, 1964; Greeno, 1966) and so can be presented here in brief.

Learning is assumed to take place by a process of trial and error. The subject selects a code for an item, and continues with that code as long as it supports retention of the item. An item becomes learned when the selected code is adequate to support retention up to the criterion of mastery in the experiment. There is a constant probability, $d$, of selecting an adequate code each time the subject samples. If an item has not yet been learned, its code will support retention for a short time, and the length of time that an unlearned item is held in short-term memory is a random variable. There is a constant probability $h$ that an unlearned item will be held in short-term memory from one trial to the next. The subject samples from the set of codes on any trial when the item is not learned and not in short-term memory.

If these assumptions are correct, we can describe the learning process as a Markov chain with four states:

- $L$: learned
- $H$: held in short-term memory
- $E$: error
- $G$: correct response due to guessing.

The initial and transition probabilities are

$$P(L, H, E, G) = [t(1 - t)(1 - g)(1 - t)g],$$

where $d$ and $h$ are the quantities defined above, $g$ is the probability of guessing correctly when an item is neither known nor held in short-term memory, and $t$ is the probability of an item’s beginning in state $L$ due to transfer of training.
Now the theoretical possibilities regarding generalization and transfer can be sharpened. According to the assumption of first-trial recognition, generalization should produce a nonzero value of $t$, applying to the $i$th item if it is not learned when the subject first learns another item with a code which includes a property of the $i$th item. However, if the subject continued on all trials to search for shared properties between the unlearned item and other learned items, then the transfer effect would include an increased value of $d$.

In an ordinary verbal concept experiment it would be difficult to separate these hypotheses because the trials on which the various items become learned cannot be specified exactly. In the present experiment, transfer of training was observed on new instances of concepts involved in a previous concept-identification task. Control items were included along with transfer items so that we could see whether transfer was more closely related to initial probabilities or transition probabilities of the learning process.

**Experimental Method**

The experiment included a number of different training conditions that correspond to different stages of learning in an ordinary verbal concept task. First, the subject learned a response to each of three sets of word stimuli. One of the sets consisted of four instances of a concept, one consisted of two instances of a concept, and the third was a single item. Then a transfer list was presented containing one new instance of each concept from the first list. With this procedure it was intended that several events would occur at the beginning of the transfer list presentation that would ordinarily occur at various times in a verbal concept task. A subject's performance on the transfer-list instance of the concept which had one training item should be like that of a subject learning a concept after he has learned the response for one of the concept instances; performance on the transfer item whose concept had two training instances should be like that of a subject who has learned the correct response for two instances of a concept in an ordinary list, etc.

Another variable was examined in the study. The instances of a concept vary in what is called the "dominance" of the concept (Underwood and Richardson, 1956). Dominance is defined in terms of the frequency with which subjects gave the concept-response to a given stimulus word in the norming association test. Thus, a high-dominant instance of a concept is one to which most subjects give the name of the concept as an associative response; a low-dominant instance is one to which only a few subjects respond with the name of the concept.

**Materials and design.** Stimulus words were selected from Underwood and Richardson's (1956) materials. Ten words were chosen as instances of each of three concept-categories: WHITE, ROUND, and SOFT. Each word had published association-frequencies with only one of these three concepts. Five words in each set elicited the concept-response relevant to this study with at least 50% frequency and are referred to as "high-dominance" stimuli. Five had only 10-20% frequency of eliciting the relevant concept-response and are referred to as "low-dominance" stimuli. In addition, nine words were selected to be used as control stimuli. Each control word had zero published association-frequency with any of the three concepts used here. There was no obvious overlap among the control words or between controls and concept-words involving concepts other than those used for selecting the concept-words.

Each subject received a training list including four instances of one concept, two instances of
a second concept, and a single instance of the remaining concept, as illustrated in Table 1. Each stimulus word was paired with one of three trigram responses: MUR, DIX, or PEL. All instances of a given concept were paired with the same response.

After this training list was learned, each subject received a test list which included a new instance of each of the three concepts as well as three control stimuli, also as illustrated in Table 1. Each of the transfer stimuli was paired with the same response as was the training instance or instances of the same concept. One control stimulus was paired with each of the three responses. Thus, each subject was tested for transfer mediated by a concept-response for which four training instances had been used, a concept-response for which two training instances had been used, and a concept-response for which one training instance had been used. For each of these tests, there was a control stimulus which was paired with the same response as the transfer stimulus.

**TABLE 1**

**EXAMPLE OF MATERIALS**

| Training list |  | Test list |
|---------------|  |           |
| Item          | Response | Concept | Item | Response | Concept |
| Knuckle       | Dix       | Round   | Globe | Dix       | Round   |
| Paste Sheep   | Mur       | White   | Atom  | Pel       | Small   |
| Freckle       | Pel       | Small   | Sulphur | Pel     | Control |
| Earthworm Tweezer |  | Grasshopper |

The dominance of training and test instances and some materials variables were varied between groups in a $2 \times 2 \times 3$ factorial design. The first factor involved two levels of dominance of the instances used in the training list. The second was the dominance of the transfer items on the test list. The third factor involved three different arrangements of materials balanced with respect to which concept had a given number of training items and which syllable was used as the response for instances of a given concept; a different set of three control words was selected for each of the three arrangements of the concept materials and responses.

**Subjects and procedure.** Eight subjects were randomly assigned to each of the 12 treatment combinations. Of the 96 subjects, 43 were students in an introductory psychology course whose participation partially fulfilled a course requirement. The remaining 53 subjects were paid $1.25 each and were recruited from introductory courses in departments other than psychology and from a group of high school seniors who were participants in a summer high school science institute.
Instructions were given for the anticipation method of paired-associate learning. Subjects were told that the stimuli would be related to one another by common descriptive adjectives or "sense-impressions." Examples were used to clarify these instructions following the procedure used by Underwood and Richardson (1956) in obtaining the normative data.

After the instructions were read the subject was asked to study three flash cards which contained the items in his training list. A response was typed on one side of each card, and on the other side was typed the word or words that were paired with that response in the training list. Three minutes were allowed for the study of these cards.

To ensure that either specific associations or concepts were learned prior to transfer, the training items were presented as a paired-associate list. Here, and later in the transfer list, items were presented by the conventional anticipation procedure with a modification intended to eliminate response learning.

Stimuli were printed on address stickers which were affixed to 4" × 6" index cards. On each card the three response alternatives were mimeographed below the stimulus. The left-to-right order of the responses was randomized on each card. Thus, the subject was always able to select from among the permitted responses, but he was unable to associate any stimulus with a response in any fixed location on the cards.

The training items were presented as a 12-item list, which included one presentation of each item in the four-item concept, two presentations of each item in the two-item concept, and four presentations of the item in the single-item concept. A deck of 36 cards was used in each training list so that three random orders of the 12-item list were presented without interruption. The subject was instructed to pronounce the stimulus word and give a response on every trial within 1 or 2 seconds after the card was presented. After he responded, the subject was told "— is correct," and the next stimulus card was presented with no appreciable delay. The items were presented until the subject gave only correct responses on three consecutive presentations of the 12-item list. No subject committed more than two errors before reaching this criterion.

After the training-list criterion had been met, the transfer list was presented. The subject was reminded that some of the words in the second list would be related to the syllables in the same way as had the words in the first list. The test lists each included six items, and a deck of 30 cards was constructed for each test list, allowing five randomizations of the list to be presented without interruption. The test list was presented until the subject gave only correct responses on five consecutive presentations of the six-item list, not including the first presentation.

Results

Figure 1 gives learning curves separately for items from concepts which had one, two, and four training items, respectively. Each curve contains data pooled over the high- and low-dominance items for both training and test. It is apparent from these curves that some transfer of training occurred, particularly on the items for which two and four training instances were given. Both the numbers of training items and the dominance of instances produced differences in the amount of transfer. These differences will be taken up in a later section.

Informal test of all-or-none transfer. We first present an informal test of the hypothesis that transfer occurred on the basis of all-or-none recognition on the first trial. This test is based on a comparison between the subset of transfer and the subset of control items obtained by eliminating all items for which there were zero errors.
These eliminations include all items for which transfer was perfect, as well as those that were learned without error during the transfer test. The remaining items, which will be called imperfect since at least one error was made on each, yield the critical data for the test.

The first-trial transfer hypothesis implies that performance should be the same for imperfect transfer items as for imperfect control items. Since imperfect transfer items have at least one error, it is hypothesized that no transfer occurred for these items. And the control items cannot have direct transfer from the training phase.

However, if opportunities for generalization occurred on trials beyond the first trial then performance would be better on the imperfect transfer items than on the imperfect controls. It also may be noted that partial transfer, such as subthreshold generalization, would result in better performance on imperfect transfer items than on imperfect controls.

The relevant data for this question are in Fig. 2, which plots the cumulative proportions of imperfect items of each type as a function of the number of errors. The data are presented separately for concepts with one, two, and four training instances because analysis of variance revealed significant differences in the rates of learning for the control items of these sets. Numerically, the differences are small, and may well be the result of response biases produced by the different numbers of stimuli paired with the responses during the initial training.
In each case in Fig. 2, the conditionalized distributions for transfer and control items appear to be nearly identical. The small differences which appear seem to be in the direction of slower learning of the imperfect transfer items than the imperfect control items.

The result given in Fig. 2 could be affected by individual differences and consequent subject-selection-biases. Fast learners might transfer most; fast learners then would be underrepresented in the curves for imperfect transfer items. A test which is not suspect on this ground was carried out. Each subject was assigned two scores. The first was the number of errors which he gave on transfer items divided by the number of transfer items on which he gave errors. The second score was the corresponding conditional mean errors per imperfect control item. The difference between these two scores was obtained for each subject. Subjects who made no errors on any of the transfer items were excluded from the analysis. The mean of the difference scores (transfer minus control) was —.074; the standard error of the mean was .172. Apparently the coincidence of the curves in Fig. 2 was not produced by a subject selection artifact.

Estimation of parameters using identifiable states. The analysis which follows makes use of statistical results which are worked out in another paper (Greeno, 1966). Briefly, the strategy of analysis is to define a set of states which can be identified directly in the data, and examine the relationship between the Markov chain with identifiable states and the Markov chain which constitutes the assumptions of the theory which we are applying. The identifiable states are

\( A: \) absorbing state, entered on the trial following the last error on an item, or on trial 1 if there are no errors on the item.
\[ R: \text{ recurrent event, in this case any error.} \]

\[ S: \text{ correct response occurring on a trial before the last error on an item.} \]

The initial and transition probabilities for the identifiable chain can be denoted

\[
P(A_1 R_t S_t) = [\pi (1 - \pi) \theta (1 - \pi) (1 - \theta)],
\]

\[
\begin{array}{ccc}
A & R & S \\
\hline
A & 1 & 0 & 0 \\
S & 0 & w & 1 - w \\
\end{array}
\]

\[
P = R \cdot u (1 - u) v (1 - u) (1 - v)
\]

It should be noted that the simplicity of the identifiable theory exhibited here is by no means guaranteed by the definitions of the identifiable states. It happens that in this instance, the assumptions of the theory imply that certain conditional probabilities in the sequences of data are constant, and these constancies permit the representation given as Eq. 2. The fact that Eqs. 1 and 2 represent equivalent theories is shown in another paper (Greeno, 1966).

The parameters as given in Eq. 2 represent the most general form of the three-state identifiable theory. Greeno (1966) shows that the assumptions of Eq. 1 imply two further restrictions, namely,

\[
v = w, \quad \theta = \frac{1 - g}{1 - gu}.
\]

Likelihood ratio tests of both these restrictions can be carried out. The test for \( v = w \) was given by Greeno and Steiner (1964, p. 330). The separate estimates for the quantities assumed to be equal and the statistical result were

\[
v = .564 \quad w = .618
\]

\[-2 \log \lambda = 2.80.
\]

The asymptotic distribution of the test statistic is \( \chi^2 \) with one degree of freedom. The probability of a value as large as 2.80 is between .05 and .10. It is worth noting that \( v \) is slightly less than \( w \). That is, before learning occurs an error is less likely following another error than following a correct response. This is somewhat surprising, since nonstationarity or individual differences would produce a set of sequences in which errors tended to follow errors more often than they should by chance. The obtained trend in the opposite direction is an indication that changes in the transition probabilities and differences in parameters among items and subjects probably had at most small effects in most of these data.
The test for \( \theta = (1 - g)/(1 - gu) \) was given by Greeno (1966). The separate estimates and the test statistic were

\[
\theta = .794 \quad \frac{1 - g}{1 - gu} = .765 \\
-2 \log \lambda = 1.95.
\]

Again, assuming the asymptotic property of a \( \chi^2 \) distribution with \( df = 1 \), the probability of a value as large as 1.95 is greater than .10.

The parameters of the identifiable theory can be estimated quite directly. Let \( N \) be the total number of sequences, and let \( N_A, N_R, \) and \( N_S \) be the numbers of sequences starting in the three identifiable states. Further, let \( n_R \) be the number of occurrences of state \( R \), and let \( n_{RA}, n_{RR} \) and \( n_{RS} \) be the numbers of transitions observed from state \( R \) to the three identifiable states. Similarly, let \( n_S \) be the number of occurrences of state \( S \) and let \( n_{SR} \) and \( n_{SS} \) be the numbers of transitions from state \( S \) to states \( R \) and \( S \), respectively. All of the \( n_i \) and \( n_{ij} \) are summed across trials and sequences within a condition for which constant parameters are assumed. Maximum likelihood estimates of the identifiable parameters are obtained by solving

\[
\hat{u}^2 = \frac{n_{RA} - N_{RR} - N_{RS}}{n_{RR} + n_{RS} + n_S} + n_{RA} = 0
\]

These estimates provide the basis for a complete quantitative analysis in terms of the theory of Eq. 1, because Eqs. 1 and 2 are observationally equivalent. Any value of the likelihood function which can be obtained with a set of values for \( u, v, \) and \( \pi \), using Eq. 2 will be equal to the likelihood of the same data obtained using Eq. 1 and the parameter values:

\[
\hat{d} = \hat{u}\theta, \\
\hat{h} = 1 - \frac{\hat{v}}{\theta}, \\
\hat{i} = 1 - \frac{1 - \hat{\pi}}{1 - gu},
\]

where

\[
\theta = \frac{1 - g}{1 - gu},
\]

as required. Therefore, Eqs. 4 are maximum likelihood estimates of the parameters of Eq. 1.
In addition to providing convenient estimates of parameters, the identifiable theory of Eq. 2 also makes the calculation of predictions extremely easy. For appropriate selections of parameters values, Eqs. 1 and 2 yield identical probabilities for any sequence of data. Therefore, they also yield identical probability distributions of summary statistics such as errors per item and trial of last error, conditional probability of an error before criterion, etc. The simplicity of the system as expressed in Eq. 2 makes derivation of several theorems a trivial matter. We take advantage of this fact now in a brief evaluation of goodness-of-fit.

It may be mentioned here that the test of goodness-of-fit to the identifiable theory does not involve the special assumption of this analysis that there is no learning while an item is in short-term memory. The identifiable theory of Eq. 2 is equivalent to any theory with states and parameters as arranged in Eq. 1 which can be obtained by inserting any probabilities of learning from states $H, E,$ and $S$ (Greeno, 1966). Suppose that we let $c$ equal the probability of learning from short-term memory. Then the present analysis assumes $c = 0$. Alternative assumptions include $c = d$ (Atkinson and Crothers, 1964), or $d = 0$ (Bernbach, 1965). Data which favor the assumption $c = 0$ over these others have been presented by Greeno (1966). However, the choice of assumption about the relative learning rates does not affect the goodness-of-fit between the theory and these data.

**Evaluation of all-or-none learning assumption.** The close coincidence between imperfect transfer items and imperfect control items informally supports the assumption that generalization to new concept instances is based on all-or-none recognition. This conclusion is also supported by formal tests of parameter invariance, which are reported below. However, these statistical tests are strongly theory-bound, and would be of little use if the all-or-none theory of learning were not applicable to these data.

As evidence relating to the all-or-none learning assumption, we present comparisons between three theorems and results from the experiment. First, let $Y$ be the number of errors in a subject-item sequence. This is the number of occurrences of state $R$ in the chain represented by Eq. 2. The probability distribution of $Y$ is

$$P(Y = i) = \pi \frac{(1 - \pi)(1 - u^{i-1}u)}{i}$$

(5)

Secondly, let $X$ be the trial number of the last error in a sequence. For a sequence with no errors, let $X = 0$. The distribution of $X$ is

$$P(X = j) = \begin{cases} \pi & j = 0, \\ (1 - \pi)u^j & j = 1, \\ (1 - \pi)(1 - u^j)(1 - uv)^{j-2}uv & j \geq 2. \end{cases}$$

(6)

Eqs. 5 and 6 represent trivial extensions of results published by Bower and Trabasso (1964).
Finally, let $W_k$ denote an error on the $k$th trial. The probability of an error on a trial before the trial of the last error is

$$P(W_k \mid X > k) = \begin{cases} \frac{\theta(1 - u)(1 - u\theta)}{(1 - u)(1 - u\theta)} & k = 1, \\ \frac{(1 - u)(1 - u\theta)(1 - uv)^{k-2}v}{(1 - u)(1 - u\theta)(1 - uv)^{k-2}v} & k \geq 2, \end{cases}$$

Equation 7 is the well-known stationarity property of all-or-none learning systems (cf. Bower, 1961; Suppes and Ginsberg, 1963). It is known now that the stationarity of this statistic depends on the assumption that performance before learning is independent from trial to trial, in addition to the assumption of all-or-none learning (Polson and Greeno, 1965). Techniques for investigating the stationarity property have not appeared previously in published articles.

To prove (7), first calculate the probability of an error on the $k$th trial. Using well-known techniques (Bower and Trabasso’s (1964) calculation is virtually identical), we obtain

$$P(W_k) = \begin{cases} (1 - \pi) \theta \quad & k = 1, \\ (1 - \pi)(1 - u\theta)(1 - uv)^{k-2}v & k \geq 2. \end{cases}$$

In the notation of Eq. 2, if the system does not absorb immediately following an error, then there will be at least one more error later on. That is

$$P(W_k \& X > k) = \begin{cases} (1 - \pi)(1 - u)\theta \quad & k = 1, \\ (1 - \pi)(1 - u\theta)(1 - uv)^{k-2}(1 - u)v & k \geq 2. \end{cases}$$

To obtain the probability of at least one error after trial $k$, sum the terms of (6) starting with the $k + 1$st.

$$P(X > k) = \sum_{j=k+1}^{\infty} P(X = j) = (1 - \pi)(1 - u\theta)uv \sum_{j=k+1}^{\infty} (1 - uv)^{j-2}$$

$$= (1 - \pi)(1 - u\theta)(1 - uv)^{k-1}.$$  

Dividing (9) by (10) gives the result.

In evaluating the adequacy of the all-or-none learning assumption, all the data were pooled in order to provide maximum power. This involves pooling over conditions where the values of $t$ are widely different; however, variations in $t$ do not affect the present analyses. First, it can be seen by studying Eqs. 4 that $\pi$ is the only identifiable parameter which depends on $t$. Secondly, a pooled estimate of $\pi$ gives an estimate of the mean value of the separate $\pi_i$’s (or a weighted average, if the different conditions have different numbers of cases). And the probability distributions of total errors and trials of last error, averaged over conditions with varying $\pi_i$, are identical to the distributions obtained by calculating from Eqs. 5 and 6 with the mean (or weighted average) of the $\pi_i$’s.
The present analyses do assume that \( u \) and \( v \) are constant across all subject-item sequences; \( u \) and \( v \) depend on \( d \) and \( h \) so the analyses assume that \( d \) and \( h \) are constant across subjects and items. The comparison already described between imperfect transfer and imperfect control items suggests that the difference between conditions probably was due to differences in \( t \), rather than in \( d \) or \( h \). Further, substantial differences in \( d \) or \( h \) between subsets of items would prevent the present theory from fitting the data satisfactorily.

The data required to calculate parameter estimates are presented in Table 2. The

**Table 2**

Frequencies of Initial States and Transitions for Identifiable-State Representation

<table>
<thead>
<tr>
<th></th>
<th>Transfer</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_A )</td>
<td>128</td>
<td>45</td>
<td>173</td>
</tr>
<tr>
<td>( N_R )</td>
<td>125</td>
<td>195</td>
<td>320</td>
</tr>
<tr>
<td>( N_S )</td>
<td>35</td>
<td>48</td>
<td>83</td>
</tr>
<tr>
<td>( n_{RA} )</td>
<td>160</td>
<td>243</td>
<td>403</td>
</tr>
<tr>
<td>( n_{RR} )</td>
<td>152</td>
<td>216</td>
<td>368</td>
</tr>
<tr>
<td>( n_{RS} )</td>
<td>125</td>
<td>155</td>
<td>280</td>
</tr>
<tr>
<td>( n_{SR} )</td>
<td>160</td>
<td>203</td>
<td>363</td>
</tr>
<tr>
<td>( n_{SS} )</td>
<td>110</td>
<td>130</td>
<td>240</td>
</tr>
</tbody>
</table>

estimated parameters of the identifiable theory, and the corresponding estimates of the theoretical parameters are

\[
\hat{u} = .388, \quad \hat{d} = .297
\]

\[
\hat{\phi} = .584, \quad \hat{h} = .237
\]

\[
\hat{\pi} = .300, \quad \hat{t} = .197
\]

The values of \( u \), \( v \), and \( \pi \) were used to calculate theoretical distributions of total errors and trial of last error according to Eqs. 5 and 6. The value of \( \theta \) substituted in those equations was

\[
\theta = \frac{1 - g}{1 - gu} = .766.
\]

The results are given in Figs. 3 and 4, where the theoretical distributions are the lines superimposed on the histograms of obtained frequencies. Goodness-of-fit was tested
Fig. 3. Predicted and observed distributions of the number of errors per item.

Fig. 4. Predicted and observed distributions of the trial of the last error per item.
by the Kolmogorov-Smirnov test. The null hypothesis would be rejected at .10 level with a maximum deviation greater than .051 between the theoretical and obtained cumulative distributions, expressed as proportions. The maximum deviation for the distribution of total errors was .031; and for the distribution of trial of last error the maximum deviation was .013.

The stationarity test was carried out by dividing each sequence of trials into two equal parts. Trials were included starting with the second trial and ending with the trial before the last error; only sequences with errors on trial 3 or later were included. If a sequence had an odd number of trials, the middle trial was considered as contributing one-half of a trial to each part of the sequence. According to Eq. 7 the number of errors divided by the number of trials should have been .464 in each part of the sequence. The obtained values were .536 for the first part and .402 for the second part. The significance of this difference was tested by assigning each subject a score equal to the proportion of errors in the first parts of his sequences minus the proportion of errors in the second parts of his sequences, as defined above. Twenty-seven subjects were excluded, not having any items with errors after trial 2. The mean difference (which should be zero, according to the theory) was $+0.112 \pm 0.081$, with 90% confidence.

A number of analyses have been presented recently, suggesting factors which might be responsible for nonstationarity. Polson et al. (1965) demonstrated two-stage learning when the likelihood of confusions among items was not constant for all the items in a list. Millward’s (1964) theoretical analysis suggests the possibility that items sometimes are partially learned by the elimination of an error response. And Atkinson and Crothers (1964) proposed that learning might proceed through two stages, with chance guessing in the initial phase and short-term memory operating to increase performance during the intermediate stage. The goodness-of-fit to the distributions in Figs. 3 and 4 indicates that the discrepancy from all-or-none learning, while significant, was not large enough to permit any meaningful evaluation of alternative factors that might have produced the discrepancy. By the same token, the all-or-none theory apparently is close enough to the truth to permit its use as a tool for analyzing the transfer effects of this experiment.

**Evaluation of Parameter Invariance.** If it is accepted that the all-or-none learning theory provides a satisfactory description of the learning process, then the all-or-none transfer hypothesis can be given an exact statistical test. Let $t_1$, $d_1$, and $h_1$ be the parameters of the learning process for transfer items, and let $t_2$, $d_2$, and $h_2$ be the parameters for control items. The hypothesis that transfer occurs on an all-or-none basis on the first trial asserts that $d_1 = d_2$ and $h_1 = h_2$, but that $t_1 > t_2$.

Further, it is reasonable to expect that $t_2 = 0$. These restrictions constitute null hypotheses that can be evaluated by likelihood ratio tests. Unfortunately, the various possible statistical tests are not independent, and the order of carrying out the tests could influence the final conclusion about the parameters. This danger is eliminated
if we carry out all possible tests and examine a matrix of results where each null hypothesis can be evaluated in combination with all the possible hypotheses about the remaining parameters.

This extensive statistical evaluation was feasible because of a computer program kindly made available by John Chandler. The program uses oscillation search to find parameter values which minimize an almost-arbitrary function. In the present case, the function minimized was $-2 \log L$, where $L$ is the likelihood function with parameters $t_1$, $t_2$, $d_1$, $d_2$, etc., with some constraint(s) imposed on the parameter values.

The results are given in Table 3. The possible constraints on $d$ and $h$ are listed in

<table>
<thead>
<tr>
<th>Subject to Various Constraints to Test Parameter Invariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d$ free</td>
</tr>
<tr>
<td>$h$ free</td>
</tr>
<tr>
<td>$t$ free</td>
</tr>
<tr>
<td>$t_1 = t_2$</td>
</tr>
<tr>
<td>$t_2 = 0$</td>
</tr>
</tbody>
</table>

Note. $4153.10$ has been subtracted from each value of $-2 \log L$ to obtain these values.

the columns of the table, and the constraints on $t$ are given in the rows. Each entry in the table is minus two times the maximum of $\log L$, minus the smallest value of this statistic in the table. (With no constraints on these parameters, $-2 \log L$ has the value $4153.10$.) A rough idea of the significance of the numbers may be taken from the fact that if $L_1$ and $L_2$ are maximum values of a likelihood function such that $L_1$ is obtained by restricting the hypothesis which leads to $L_2$, then $-2 (\log L_1 - \log L_2)$ is distributed asymptotically as $\chi^2$ with degrees of freedom determined by the number of constraints used in obtaining $L_1$. For example, in Table 3, we may select an assumption about $t$. This determines a row in the table. Now, assuming that the selected constraint is correct, we can compare the left-most figure in the row with the right-most figure in the row. If $d_1 = d_2$ and $h_1 = h_2$ are both true, then the difference

$^2$ The computing routine is titled Subroutine Stepit and is available in FORTRAN. It may be obtained from the Quantum Chemistry Program Exchange, Indiana University; request program QCPE 66.
should be approximately distributed as $\chi^2$ with two degrees of freedom. The parenthesized figures in Table 3 are the $\chi^2$ degrees of freedom relative to the hypothesis of no constraints.

The statistical results in Table 3 indicate rather clearly what went on in the experiment. All of the test statistics for hypotheses which include $t_1 = t_2$ are large and highly significant relative to corresponding hypotheses that do not include the hypothesis $t_1 = t_2$. All of the test statistics for other hypotheses are small and nonsignificant. The statistical results are consistent with the following conclusions: (1) there was a transfer effect in this experiment; (2) virtually all of the transfer was produced by all-or-none recognition on the first presentation of new items; and (3) the a priori expectation that $t_2 = 0$ was satisfactory.

**Statistical evaluation of transfer effects.** The preceding analysis has shown that there was a difference between transfer and control items in this experiment. Now we ask whether the different transfer conditions produced different amounts of transfer. There were three treatment factors: number of training examples, dominance of training examples, and dominance of test examples. The statistical analysis was carried out here in the same way as above, with the simplifying restrictions that seem warranted on the basis of the preceding analysis. That is, it is assumed that $t = 0$ for control items and that a single value of $d$ and a single value of $h$ apply to all items. With these assumptions, maximum likelihood estimates of $t$ were obtained for the various transfer conditions, subject to various constraints. The results are in Table 4.

<p>| TABLE 4 |
|------------------|------------------|
| <strong>VALUES OF $-2 \log L$, WITH $L$ MAXIMIZED</strong> |
| <strong>SUBJECT TO VARIOUS CONSTRAINTS TO TEST EXPERIMENTAL EFFECTS</strong> |
| No. of examples | No. of examples |</p>
<table>
<thead>
<tr>
<th>Varying</th>
<th>Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train. dom. varying Test dom. varying</td>
<td>0.00</td>
</tr>
<tr>
<td>Train. dom. varying Test dom. constant</td>
<td>15.28 (1)</td>
</tr>
<tr>
<td>Train. dom. constant Test dom. varying</td>
<td>36.94 (1)</td>
</tr>
<tr>
<td>Train. dom. constant Test dom. constant</td>
<td>45.47 (2)</td>
</tr>
</tbody>
</table>

*Note.* 4087.70 has been subtracted from each value of $-2 (\max \log L)$ to obtain these entries.
presented in the same way as in Table 3. Holding $t$ constant for either test dominance conditions or training dominance conditions involves a single restriction of the parameters. Holding $t$ constant for the different numbers of examples involves two restrictions. The implication of Table 4 is that $t$ must be considered to have varied as a function of all three of the experimental conditions.

**Parametric Analysis of Transfer Effects.**

It is established that $t$ varied among the different experimental conditions. The present section presents the results of an attempt to account for the variation in $t$.

With just 24 transfer items in each of the twelve conditions, a reliable estimate of each separate $t$ cannot be obtained, especially because all the available information about $t$ comes from the proportion of items with zero errors. However, an impression of the magnitude of variation can be obtained from Table 5, where data are presented on an analogy with the “main effects” of analysis of variance. In the first four lines of Table 5 are estimates of $t$ for the four dominance conditions, averaged over the three numbers of training instances. In the last three lines are estimates of $t$ for the three numbers of training examples, averaged over the dominance conditions. The column labeled “Empirical $t$” has values estimated separately for the various conditions specified in the table. The values labeled “Theoretical $t$” were obtained by an analysis which is now described.

<table>
<thead>
<tr>
<th>Dominance</th>
<th>Empirical</th>
<th>Theoretical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>.616</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>.409</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>.233</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>.185</td>
</tr>
<tr>
<td>No. of Training Examples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.557</td>
<td>.552</td>
</tr>
<tr>
<td>2</td>
<td>.365</td>
<td>.372</td>
</tr>
<tr>
<td>1</td>
<td>.161</td>
<td>.222</td>
</tr>
</tbody>
</table>
According to the general considerations offered in Section 1, the amount of transfer in a given condition should depend on two factors, (a) the probability that the subject's code for one or more training items includes an associative component which will permit transfer to the test item, and (b) the probability that the subject recognizes the shared component when the test item is presented. The evidence presented in Section 4 relates to the nature of the transfer effect when it occurs. The two transfer factors apparently determine the value of $t$ and have no effect on the other parameters of the learning process.

One simple analysis can be obtained by assuming that $t$ is the product of two probabilities,

$$t_{ijk} = A_{ij} b_k, \quad (i = 1, 2, 4; j = H, L; k = H, L), \quad (11)$$

where $i$ is the number of training examples, $j$ is the dominance of the training examples, and $k$ is the dominance of the test example. $A_{ij}$ is the probability that the subject will code at least one of the $i$ training examples in a way that includes the associative element that is shared by the test example. And $b_k$ is the probability that the subject will recognize the shared component in the test, given that the initial learning did include the shared component in at least one code. The strong assumption of Eq. 11 is that the probability of recognition does not vary with the number of training examples whose codes include the concept-association as a component. If Eq. 11 is correct, then we can consider the concept-association to be available to support transfer or not, on an all-or-none basis.

Now consider the value of $A_{ij}$. Assume that there is a probability, $a_j$, varying with dominance but constant otherwise, that the subject will select a code for memory that makes the concept-association available when he learns an item. Since there are $i$ items, the probability of including the concept-association for at least one of them is

$$A_{ij} = 1 - (1 - a_j)^i. \quad (12)$$

Equation 12 involves the strong assumption that the probability of coding a given training item with the concept association was independent of the number of training items which were present on the stimulus card during training. It presupposes an idealization with the subject scanning the words on a stimulus card, memorizing the word-syllable pairs one by one. Each time a pair becomes learned there is a chance that its code makes the concept-association available. There are no assumptions involved here about whether the code transfers to the other training items on the card. Presumably it would with high probability, but that would not affect the present analysis which assumes that transfer is independent of the number of examples coded to include the concept-association.

The relevant data for analyzing transfer are just the proportions of items with zero errors in the various conditions. The appropriate measure of goodness-of-fit is $\chi^2$. 
Chandler's computing routine was used to find values of $a_H$, $a_L$, $b_H$, and $b_L$ giving a minimum value of $\chi^2$. The obtained estimates were

$$a_H = .387, \quad a_L = .147,$$

$$b_H = 1.000, \quad b_L = .600.$$

The results are in Table 6. As can be seen there, it was necessary to pool the two- and four-example frequencies when data for both training and test dominance were

<table>
<thead>
<tr>
<th>Condition</th>
<th>Training dominance</th>
<th>Test dominance</th>
<th>No. of examples</th>
<th>$t$</th>
<th>Predicted frequency</th>
<th>Obtained frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
<td>4</td>
<td>.858</td>
<td>21.05</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>2</td>
<td>.624</td>
<td>16.15</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>1</td>
<td>.387</td>
<td>11.20</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>4</td>
<td>.566</td>
<td>14.95</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>2</td>
<td>.412</td>
<td>11.72</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>1</td>
<td>.255</td>
<td>8.45</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>4</td>
<td>.472</td>
<td>12.97</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>2</td>
<td>.273</td>
<td>8.83</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>1</td>
<td>.147</td>
<td>6.21</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>4</td>
<td>.311</td>
<td>9.63</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>2</td>
<td>.180</td>
<td>6.89</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>1</td>
<td>.097</td>
<td>5.16</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

high, in order to collect a large enough expected frequency. The observed frequencies do not deviate significantly from the predictions: $\chi^2$ is 13.75 which has $p > .05$ with 7 df. It appears that the assumed independence between items may not have held in the case of low-dominance training items, but the data are not sufficiently powerful at this level to permit more detailed analysis. The average values of $t$ that were calculated from these assumptions are presented in Table 5, and seem to be satisfactorily close to the estimated averages.

The analysis in terms of the $a_i$ and $b_j$ parameters serves a useful purpose here, since it clarifies the quantitative relationships among the experimental effects. Further, the assumptions seem close enough to the truth to serve as the basis for future investigations of verbal concept acquisition. The results of the present experiment provide firm evidence about the process of generalization and transfer that occurs after a
verbal concept has been recognized by a subject. Questions involving the possible
independence of items and the ways in which subjects code and compare items in
memory can be evaluated best in relation to direct experimental tests of alternative
hypotheses about those processes.

REFERENCES

Atkinson, R. C., and Crothers, E. J. A comparison of paired-associate learning models having
different acquisition and retention axioms. Journal of Mathematical Psychology, 1964, 1, 285-
315.

Atkinson, R. C., and Estes, W. K. Stimulus sampling theory. In R. D. Luce, R. R. Bush, and
Pp. 121-268.

Bernbach, H. A. A forgetting model for paired-associate learning. Journal of Mathematical
Psychology, 1965, 2, 128-144.

Bower, G. H. Application of a model to paired-associate learning. Psychometrika, 1961, 26,
255-280.

Bower, G. H. An association model for response and training variables in paired-associate


Estes, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush

Estes, W. K. Learning theory and the new "mental chemistry." Psychological Review, 1960, 67,
207-223.

Estes, W. K. All-or-none processes in learning and retention. American Psychologist, 1964,
19, 16-25.

Friedman, M. P. Application of statistical learning theory to component-vs.-compound dis-

Friedman, M. P., and Gelfand, H. Transfer effects in discrimination learning. Journal of
Mathematical Psychology, 1964, 1, 204-214.

Greeno, J. G. Paired-associate learning with short term retention: mathematical analysis and

Greeno, J. G., and Steiner, T. E. Markovian processes with identifiable states: general con-

Millward, R. An all-or-none model with elimination of incorrect responses. Journal of Ma-
themtical Psychology, 1964, 1, 392-404.


Polson, M. C., Restle, R., and Polson, P. G. Association and discrimination in paired-associates


Restle, F. The selection of strategies in cue learning. Psychological Review, 1962, 69,
329-343.

Restle, F. Sources of difficulty in learning paired associates. In R. C. Atkinson (Ed.), Studies in
RESTLE, F. A cognitive interpretation of intensity effects in stimulus generalization. Psychological Review, 1964, 71, 514-516. (b)


SUPPES, P., AND GINSBERG, R. A fundamental property of all-or-none models, binomial distribution of responses prior to conditioning, with application to concept formation in children. Psychological Review, 1963, 70, 139-161.


UNDERWOOD, B. J., AND RICHARDSON, J. Some verbal materials for the study of concept formation. Psychological Bulletin, 1956, 53, 84-95. (b)

Received: September 17, 1964