REINFORCEMENT IN HUMAN LEARNING

BY

W. K. ESTES

TECHNICAL REPORT NO. 125

DECEMBER 20, 1967

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

Reinforcement in Human Learning

by

W. K. Estes

Technical Report No. 125
December 20 1967

Psychology Series

Institute for Mathematical Studies in the Social Sciences
Stanford University
Stanford, California

Preparation of this article was supported by Contract Nonr 225(73) between the Office of Naval Research and Stanford University. Experimental research reported herein was supported in part by a grant from the National Science Foundation.

I am indebted to Alexander M. Buchwald for his painstaking criticisms of a preliminary draft of the article.

Introduction

Although the interpretation of reinforcement in animal learning has been a focus of theoretical controversy for many decades, the corresponding issue as it arises in experimentation on human learning has been strangely quiescent for a very long period. Perhaps the reason is in part that much of the research on human learning, at least in this country, has been conducted by investigators in the functionalist tradition. Although functionalism entails no necessary commitment to any particular theory of reinforcement, its representatives seem usually to have proceeded on the assumption that the nature of reward is captured in essentials by Thorndike's law of effect, according to which the action of reward is a direct and automatic strengthening of the stimulus-response association.

Following Thorndike's systematic studies of "aftereffects" (1931, 1935), the problem of reward in human learning received little explicit attention until the recent burgeoning of research on verbal behavior by means of operant conditioning techniques. The wave of these studies, which followed Greenspoon's original experiment (1955) appeared uniformly to support an operant conditioning interpretation essentially equivalent to the law of effect (see, e.g., Salzinger, 1959). It appeared that suitably programmed rewards controlled the occurrence of verbal behaviors in a manner predictable from analogous studies of operant behavior in animals, and, in particular, that effects of rewards were independent of the subjects' awareness of relationships or contingencies between their responses and reinforcing operations.

In a major review of the literature on verbal rewards and punishments Postman and Sassenrath (1961) concluded (1) that the law of effect was on the whole well supported by a variety of evidence; (2) that of the various types of evidence that other investigators had thought to bring the principle in question, only that having to do with the necessity of awareness of response-reinforcement contingencies on the part of the learner was truly germane to the issue; and (3) that, owing to methodological and logical difficulties having to do with criteria of awareness, the studies purporting to relate effectiveness of rewards to the subject's awareness of contingencies cannot provide a source of critical evidence regarding tenability of the law of effect.

In his own theoretical writings, Thorndike (1931) recognized, and attempted to refute, an alternative type of theory in which rewards and punishments might function as determiners of choices but without any direct influence on learning. According to this conception, the learning which occurs in a standard human learning experiment would be a matter of the subject's acquiring information concerning relationships between stimuli and responses in the material to which he is exposed and relationships between these and rewarding or punishing outcomes. The learning would simply be a function of such variables as contiguity, frequency, and recency. The values of rewards and punishments would be relevant only in so far as they determined the subject's choice of responses on a given trial once he had recalled the relevant events that had occurred on the preceding trials of an experiment. This alternative viewpoint has received little attention in the literature

on human learning, perhaps partly because Thorndike did not favor it himself and perhaps more importantly because no explicit and detailed theory of this type has been formulated.

More recently, two research developments have lent new interest to the desirability of formulating a theory of the second type. Firstly, with regard to the operant conditioning situation, studies by Dulany (1962) and by Spielberger and his associates (1962, 1966) have cast considerable doubt upon the conclusion that verbal response tendencies are systematically modified by the effects of rewards when the subjects are unaware of relationships between the rewarding events and their responses. With regard to the more classical types of human learning experiments, involving the paired-associate or Thorndikian paradigms, a series of studies by myself and associates (Keller, et al. 1965; Estes, 1966; Humphreys, Allen, and Estes, in press) have provided considerable evidence that the function of reward in these situations can be experimentally decomposed into what may be termed informational and motivational components.

In the remainder of this chapter I propose, firstly, to review a sample of earlier findings from the rather extensive series of studies done in my laboratory leading to an interpretation of reinforcement in human learning that does not assume the law of effect; secondly, to outline the theory that has taken form; and, thirdly, to present one of the relatively novel experiments we have contrived to put this formulation to an exacting test.

Information Versus Effect

When a reward is presented following a response, there are at least two distinct ways in which the reward might affect learning of the response. Firstly, there is the possibility that the reward exerts a direct strengthening effect upon the associative connection between the response and the stimulus which evoked it, this being of course the conventional law of effect interpretation. Secondly, there is the possibility that the subject simply learns the relationships between the preceding stimulus and response and the reward as a result of having experienced them in temporal contiguity. A long term experimental program in my laboratory has been directed toward the problem of separating these two possible functions of reward.

The first set of data relevant to this issue comes from two groups included in a larger experiment that was conducted by a group of graduate students in my advanced laboratory methods course at Indiana University in 1957. The purpose was to ascertain the extent to which learning of a paired-associate item could be influenced by reward in the sense of an indication of correctness of response when opportunities for acquiring information about stimulus-response-reward relationships were strictly equated. For brevity let us denote the two conditions as being the Effect and the Information conditions. Twenty-four college student subjects were run under each condition, learning 12-item lists of double letter, double digit pairs. Subjects in the Effect condition received first a cycle through the list on which each of the stimulus-response pairs was presented once, then a series of cycles under a

conventional anticipation procedure, continuing until the subject met a criterion of all 12 responses correct through a single cycle.

Under the anticipation procedure of the Effect condition, each trial began with a 3-sec. exposure of the stimulus member of an item during which the subject attempted to give the correct response from memory, then a 2-sec. blank interval, then a 3-sec. presentation of the stimulus together with its correct response. A 60-sec. rest interval intervened between successive cycles through the list and, of course, the order of items was randomized anew for each cycle.

For the Information group, odd-numbered cycles through the list involved only paired presentations of the stimulus-response members of the items; and even numbered cycles comprised recall test trials on which the stimulus members were presented alone, with the subject attempting to give the response from memory but with no feedback of any kind from the experimenter indicating correctness or incorrectness of the response. Under this procedure, on any one trial of the paired presentation cycle the subject received the 3-sec. presentation of the stimulus response pair, then a 2-sec. blank interval, then a 3-sec. presentation of the next scheduled pair, and so on; a 30-sec. rest interval intervened between the end of this cycle and the following test cycle. On a test trial each stimulus member appeared alone for 3 sec., during which the subject responded, with 2 sec. between successive trials and a 30-sec. rest at the end of a cycle before initiation of the following paired presentation cycle. Thus, within a margin of error of 2 sec. per cycle on the average, the

interval between a paired presentation of any item and the next subsequent recall test on that item was strictly equated for the Effect and Information conditions. The distinctive difference between the procedures was that under the Effect condition subjects received immediate reward, in the sense of an indication of correctness, following every correct response; under the Information condition a subject's correct response on a recall test could have received no reward until that item came up on the subsequent paired presentation cycle, in general more than a minute later and after a large number of intervening items had been presented. Whereas both groups received exactly the same amount of information concerning correct stimulus-response relationships between successive recall tests, the Effect group should have had an advantage if an immediate indication of correctness exerts a strengthening effect on stimulus response association in the manner assumed by the classical law of effect.

The principal results are presented in Table 1 in terms of mean numbers of trials and mean numbers of errors to criterion for the two conditions. Clearly the Effect condition yielded no advantage, and in fact slower learning according to both measures, though in neither case was the difference statistically reliable. The lack of any advantage for the group receiving immediate knowledge of results is evidently not peculiar to our conditions, for it has subsequently been replicated several times (e.g., Battig and Brackett, 1961). This finding, rather unexpected from the viewpoint of law of effect theory, could perhaps be explained in various ways, but the simplest interpretation would

Table 1

Trials and Errors to Criterion under Information and Effect Conditions

Condition	Trials		Errors	
	M	SE_M	M	SE_M
Information	7.71	0.64	37.79	3.81
Effect	8.96	0.48	45.83	3.05

BLANK PAGE

seem to be that learning of the correct stimulus-response relationships in this situation depends simply upon contiguous experience of the corresponding events and not upon rewarding consequences following correct responses.

Having failed to detect any strengthening affect of immediate knowledge of results when information presented was held constant across conditions, we turned next to the mirror image, that is, a comparison with reward value held constant while information was allowed to differ. In a study by Keller, Cole, Burke, and Estes (1965), subjects learned a 25-item paired-associate list, all items having different stimulus members but there being only two response alternatives, one or the other of which was correct for each item. On each trial of the experiment the stimulus member of an item appeared on a screen, then the subject chose one of the two possible responses and, following this, received the reward assigned to that stimulus response combination. Rewards were given in "points," the numerical values of which were presented visually immediately following the response; the number of points received over a series of trials was directly related to the subject's monetary payoff for participating in the experiment.

For each stimulus in the list a pair of reward values, for example 1 point vs 8, 2 points vs 4, was assigned to the two possible responses and the assignment did not change for the given subject throughout the experiment. However, the conditions of presentation of the rewards differed for two groups, one group being run under what we shall call the Correction and the other the Noncorrection procedure. The

Noncorrection condition corresponded essentially to that of the Thorndikian trial and error experiments; that is, on each trial a subject made his choice of responses following presentation of the stimulus and then was shown the reward value given for that response. Under the Correction condition, once the subject had made his choice on each trial, both assigned reward values were shown, the one which the subject received for the response he had made and also the one associated with the response not made on that trial. Thus, the rewards actually received for each stimulus-response combination were exactly the same for the two groups, but the amount of information given per trial concerning stimulus-response-reward relationships was twice as large for the Correction as for the Noncorrection condition.

With an error on any item defined as a choice of the response carrying the lower reward value, mean errors per item to a criterion of five consecutive correct responses were compared for items having different reward combinations within each of the training conditions. For the Noncorrection condition, the overall picture was essentially as one would anticipate on the basis of classical conceptions of reward, the rate of learning being higher the greater the difference between the reward values assigned to the two response alternatives of an item; and a similar ordering appeared with respect to mean trial of the last error. In each case the variation over reward combinations was highly significant, the mean value for the item with the largest reward differential (1 point vs.8) being less than half that for the slowest learned item (2 vs.4). For the Correction condition, in contrast,

there was no significant variation in either of these statistics over items with different reward combinations and no consistent ordering of items with respect to reward differentials. For example, the fastest learned item was 4 vs.8 with 1.84 errors and the slowest 1 vs.4 with 3.07 errors, with such extreme differential items as 1 vs.8 and 1 vs.2 falling in between.

The first of these sets of findings fits in with the notion that a larger difference between the reward values for an item leads to a drawing apart of associative strength for the two response alternatives simply because of the different increments in strength produced by the two rewards. However, the correction procedure involved the same reward contingencies, yet led to no significant variation of learning rate with reward differentials, thus casting doubt upon that interpretation. When full information was given on each trial for each item, the relationship between learning rate and reward differential disappeared.

If learning rate were a function purely of information received by the subject per trial, we should expect the rate to be approximately twice as great for the Correction as for the Noncorrection condition, since in the former information concerning both of the response-reward relations of an item was given on each trial but in the Noncorrection condition only one of the two was available to the subject on any one trial. Interestingly, the rate of learning as measured by the slope constant of the learning curve was almost exactly twice as large for the Correction as for the Noncorrection condition (.21 vs. .10). If learning is actually a function only of information transmitted, then

the variation among observed learning curves for the Noncorrection condition must be attributed simply to a performance difference, expected on the ground that, in the case of partially learned items, subjects will tend to make fewer errors if the reward differentials are large. When only one alternative has been learned for a given item, that choice will tend to be made if the learned reward is large; that choice will tend to be avoided if the learned reward is small.

Additional data of considerable value both for evaluating the classical interpretation and for pointing the way to desirable modifications are available in the response latencies of the study by Keller, et al. Considering mean latencies over the entire series as a function of reward combinations, for the Correction condition there were only slight differences, and the only noticeable systematic trend was a slight tendency for latencies to decrease as the sum of the reward values per item increased. For the Noncorrection condition, however, a much steeper function emerged, with mean latency decreasing as the higher reward value of a pair increased and, when the higher value was held constant, decreasing as the sum of the reward values increased.

Some points of special interest with respect to the Noncorrection data are brought out in Figure 1, in which mean correct response latencies are plotted for the first block of 10 trials in the upper two panels and for the final block in the lower two panels. On the left-hand side, latencies are plotted as a function of reward differential per item. A significant, though modest correlation is apparent, but the correlation

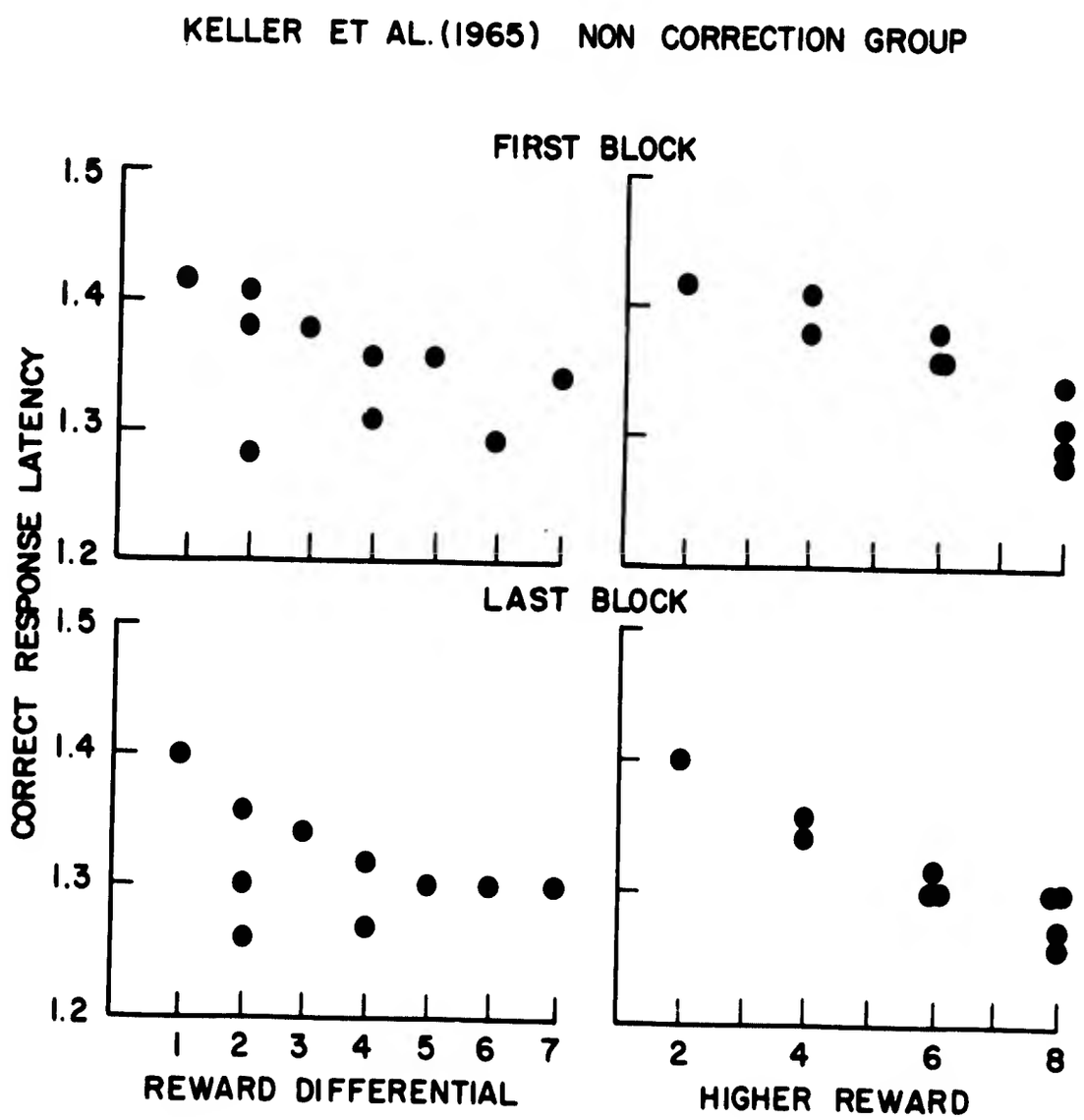

Figure 1. Mean correct response latency as a function of reward differential (left-hand panels) and of value of the higher reward per item (right-hand panels) at early and late stages of learning.

BLANK PAGE

decreases somewhat from the first block to the last; in fact, if the point for the reward differential of 1 (the 2 vs. 1 item) were deleted, the correlation would be virtually 0 in the last block.

In the right-hand side, latencies are plotted as a function of the value of the higher reward per item. The two significant findings which emerge from the latter are that the correlations of latency with the independent variable are higher in both instances on the right than on the left, and that the correlation increases from the first block to the last, the function being virtually linear in the final block.

Evidently, the higher the reward value a subject has learned to expect following a given response, the more probable is that response to occur and the lower is its latency on the average when it does occur.

Still more direct demonstrations of the establishment of learned associations between originally neutral stimuli and rewards, and of the effects of this learning upon performance in new situations, is provided by several recent studies conducted with children. For example, Nunnally, Duchnowski, and Parker (1965) gave second-grade children experience with an apparatus constructed somewhat like a roulette wheel except that the wheel was bordered with nonsense syllables rather than numbers. When the wheel stopped with the pointer on a particular one of the syllables, the positive stimulus, the subject received a two-cent reward; when it stopped on a second syllable, the negative stimulus, the subject was punished by loss of one cent; and when the wheel stopped on the third syllable the outcome was neutral, that is, neither gain nor

loss. After a series of training trials, tests were given in a "treasure hunt" situation in which the subject was confronted with a set of white boxes, identical except that each had one of the nonsense syllables printed on the top, and was given a series of opportunities to find the box which contained a twenty-five cent reward. The subjects exhibited a significant tendency to select boxes labelled with the previously positive stimulus as compared to the previously negative or neutral stimuli. Further, when in another test the subjects were simply allowed to look at cards containing the various nonsense syllables in a view box with their eye movements being photographed, they exhibited a greater tendency to look at the positive than the neutral stimulus and longer at the neutral than the negative stimulus. Similar results concerning "looking time" were obtained in a related series of experiments reported by Nunnally, Stevens, and Hall (1965).

In a study of somewhat different design, Witryol, Lowden, and Fagan (1967) gave children training intended to establish associations between stimulus dimensions and reward values. During a training series the child made a choice and then received a reward. When two stimuli from one dimension, say color, were presented, the child received a high reward regardless of his choice, and when stimuli from the other dimension were presented the child received a low reward regardless of his choice. Following this training, the children were given a series of trials on a discrimination problem in which one of the formerly presented dimensions was relevant and the other irrelevant with, now, a light serving as the reinforcing stimulus indicating correctness

of response. Discrimination learning was more rapid when the formerly high reward dimension was relevant, and this finding, together with other aspects of the data, indicated that the children had developed an increased tendency to attend to the dimension associated with high reward.

Provisional Interpretation of Functions of Reward

The findings just reviewed are representative of the series of studies which appears to be converging in a rather direct way upon a particular set of assumptions regarding the functions of reward. The ideas involved are applicable to animal as well as human learning and can be cast in mathematical form up to a point (see Estes, 1967; Humphreys, Allen and Estes, 1968 [in press]) but for present purposes it will suffice to develop them informally in the context of standard human learning situations.

Any one trial of a simple trial and error learning experiment, such as that of the study by Keller, et al., involves presentation of a stimulus, a choice of response by the subject, and finally presentation of a rewarding outcome. For brevity we shall denote these three components by S_i, R_j, and O_k. The principal assumption to be made concerning learning is that associations between any two of these components may be formed simply as a function of contiguous experiences of them. Thus on a trial in which a stimulus S is followed by a response R_1 and outcome O_1, associations of the form $S\text{-}R_1$, $S\text{-}O_1$, and $R_1\text{-}O_1$ may be established. It is assumed that learning in this sense occurs on an all-or-none basis, with a constant probability over trials that an association

will form upon the contiguous occurrence of the appropriate members. The term association is being used simply in the sense of memory storage such that, in the case of an S-R association, the presentation of S evokes a representation of R, in the case of an S-O association the presentation of S evokes a representation of O, and so on. This learning is assumed to occur purely as a function of contiguity and independently of reward values of outcomes.

One aspect of this last assumption requires clarification. What is proposed is that once two stimuli are sampled, or perceived, contiguously there is some probability that an association will form, this probability being independent of associated reward values. It is not implied, however, that the probabilities of sampling particular stimuli may not change over a series of trials. In fact, application of the present theory to the learning of perceptual, or orienting responses, leads to the prediction that subjects' selective sampling of stimuli may be systematically modified by a learning series in which different cues are associated with different rewards.

The principal function of rewards in this theory is to modify response selection in any given state of learning. If learning has occurred with respect to a particular item, then when the item is presented the stimulus members are scanned and their learned associations are activated, that is brought into the subject's immediate memory. It is assumed that recall of a rewarding or punishing outcome carries what may be termed facilitatory or inhibitory feedback. The effect of modulation by feedback is that when the scanning process moves from a

recalled response with anticipated outcome of higher to one of lower reward value, feedback becomes inhibitory and the latter response does not occur overtly. When the process moves from a lower to a higher recalled reward, the latter response-reward association carries facilitatory feedback and there is some probability, directly related to reward value, that the response will occur overtly.

Thus, when an item is in the fully learned state, that is both alternative responses and their associated rewards have been learned, then, upon each presentation of the stimulus, the correct response is chosen and the latency of the response is shorter the higher the associated reward value. When an item is completely unlearned, no representations of previous responses or rewards are brought into memory when the stimulus is presented and the subject simply must choose at random between the available alternatives. When an item is partially learned, so that upon occurrence of the stimulus only one of the associated response-reward combinations can be recalled, the outcome is not fully determinate. Presumably the feedback associated with the response whose reward value is unknown would be equal to the average of that generated by choices of responses with unknown rewards during the preceding trials of the experiment. Since this value would fluctuate somewhat from trial to trial, all that can be assumed is that for any given known reward value there is some probability that it would carry higher positive feedback than that of the unknown reward value associated with an alternative response to the same stimulus, and this probability would be higher the higher the value of the known reward.

Even though our assumption be correct that associative learning of a stimulus-response relationship depends solely upon the contiguous experience of the two members by the learner, it must be recognized that other factors influence performance under various test conditions. There is considerable reason to believe that for recognition of learned stimulus-response relationships, the contiguity assumption is for all practical purposes sufficient as well as necessary, (see e.g., Estes and Da Polito, 1967). By contrast, any type of recall test, involving determination of the subject's tendency to give the response upon presentation of the stimulus alone, is sensitive to variables influencing covert or overt rehearsal. In the study by Estes and Da Polito for example, introduction of an instructional condition intended to reduce the subject's tendency to rehearse stimulus response relationships during training had virtually no effect upon subsequent recognition performance but was severely detrimental to subsequent recall.

Within stimulus sampling theory (Estes, 1959) there are several reasons for expecting this dependence of recall performance upon rehearsal. Firstly, the repetition involved provides repeated opportunities for elements of the stimulus to become associated with the response. Secondly, during rehearsal the stimulus member occurs out of the context of the training trial, which might be expected to promote transfer to later tests in which the stimulus appears in different contexts. Thirdly, in rehearsal the stimulus and the response members of the association will frequently be brought into closer temporal contiguity than was the case under conditions of a training trial.

The pertinence of these observations for the present discussion is that in adult human subjects, at least, it must be expected that motivational variables and, in particular, rewards and punishments will influence rehearsal. The effects may be expected to be manifest in at least two ways. Firstly, if one compares conditions for different groups of subjects which differ with respect to reinforcing conditions, we must expect an overall difference in performance, other things equal, in favor of the group receiving the greater rewards or punishments, since either would lead to increased tendencies to rehearse and thus to an improvement in performance which would be nonspecific with respect to items. Secondly, in experiments in which reinforcing conditions vary from item to item, we should expect greater tendencies to rehearse on items involving larger reward values. However, it should be noted that the essential relationship here is between the stimulus members of the item and the associated rewards, not the specific relations between responses and rewards. Thus, for example, in the Correction condition of the Keller, et al. (1965) study, an overall tendency was manifest for response latencies to decrease as a function of the sum of the reward values associated with the stimulus member of an item. This effect was slight, however, in conformity with a large amount of experience indicating that when adult human subjects participate in experiments under instructions to learn, sufficient rehearsal occurs so that differences attributable to this factor are very hard to demonstrate. When, however, the learning task is masked in some fashion from the subjects, as in many studies of operant conditioning of verbal

behavior, of incidental learning, or of learning without awareness, differential rehearsal becomes a major variable and must be taken into account.

An Experimental Test of the Theory

The distinctive feature of the interpretation just outlined is that a reward is assumed to exert no direct effect upon associative connections at the time of its occurrence, but to exert its effects upon performance only upon later occasions after the relationships between stimuli and reward have been learned. That is, rewards do not influence learning; but rather information concerning rewards must be acquired before rewards can influence performance. This conception accounts for the experimental findings mentioned in preceding sections of this paper, but we should like to see the central assumptions put to a more direct test. An example of the type of experiment which seems to be called for is the following, hitherto unpublished study conducted at Indiana University in 1961.

In all traditional trial and error learning experiments, rewards are presented immediately following stimulus response occurrences, and thus the opportunity for any strengthening effect of the reward upon the stimulus-response association and the opportunity for the learning of relationships between the stimulus-response combination and reward are inextricably confounded. What I wished to accomplish in this study was to separate these two possible aspects by contriving a situation in which the subject could learn relationships between stimulus-response combinations and reward values without any possibility of a direct

strengthening effect of the latter upon the former. This end was accomplished by the following procedure.

The task was a modification of the verbal discrimination experiment which has been used previously by Bower (1962), Estes (1966), and others. During the training phase subjects were given repeated cycles through a set of pairs of stimulus cards. Reward values were assigned to the cards, and on each trial the subject was to choose the card of the pair which he thought carried the larger reward value. Up to this point the procedure is the same as that of previous studies of two-choice, differential reward learning with the Noncorrection procedure (Estes, 1966; Keller, Cole, Burke and Estes, 1965). However, in the present experiment the subject's task was complicated in that, in order to receive the reward, he had not only to choose the correct card of a pair but also to guess correctly the precise reward value assigned to that card. In all cases the incorrect card of a pair had reward value 0. For some pairs some single reward value was assigned to the other card and did not change over the series of trials. For the items of the type just described, belonging to what we shall term the Uniform condition, whenever the subject made a choice of a card together with his guess at the reward value, he was then shown the assigned reward value and, if he had both chosen and guessed correctly, was given that reward.

For the set of items belonging to what we shall term the Random condition, one card of a pair carried 0 reward and the other either of two values, for example 1 or 2 units. For these items, upon correct choices the reward value displayed by the experimenter at the end of

the trial was chosen randomly from the two assigned values; hence, when the subject had learned the values assigned to the card he would on the average guess the value correctly and thus receive the reward on only half the trials.

For a third set of items, belonging to what we shall term the "Never-right" condition, the assignment of rewards to the two cards of a pair was identical to that of a corresponding item in the Random condition. But the procedure differed in the following way. Whenever the subject chose the correct card of a pair and guessed a reward value belonging to the assigned set, for example, 1 if the pair was 1 or 2, the reward value displayed by the experimenter was always the one not guessed by the subject. Thus on these items, the subject could learn the pair of reward values assigned to the correct card of a pair but he could never actually receive any reward for his choice. Since the three types of items were mixed together and given in new random order on each cycle during the training phase of the experiment, it was expected that the subjects would be able to learn the various stimulus-response-reward relationships but would not be able to perceive the overall structure of the experiment and, in particular, would not be aware that they never actually received rewards on some particular items.

After the subjects had learned the correct responses to all items, in the sense of always choosing the correct member of each pair and always guessing a reward value which belonged to the assigned subset, test trials were given in which the non-zero cards from the training pairs were recombined in various ways. By means of these tests it was hoped

that we could determine whether a subject's tendency to choose one card over another in a new combination would be determined by the rewards actually received for choices of these cards during previous training, or simply by the reward values previously assigned to the cards, which the subject had had opportunity to learn even though he might never have received the given rewards. The following is a more detailed exposition of the methodology of the experiment.

Method

Subjects and materials

The subjects were 40 undergraduate students at Indiana University who received $2.00 each for participation in the experiment. The stimuli were presented on 3 x 5-in. cards, each of which contained a nonsense syllable, that is a consonant-vowel-consonant combination printed in capital letters, on the face, and the assigned reward values on the back. During the experiment the cards were kept in a rack behind a shield and were presented in pairs by the experimenter, the subject seeing the stimulus sides of the cards but never being shown the sides of the cards containing the reward values. Behind the shield was also a bank of electric counters. A set of 16 nonsense syllables was used as stimuli, with a different random assignment of syllables to reward values for each subgroup of 10 subjects.

Stimulus-reward conditions

The 16 stimulus cards used for a given subgroup of subjects were assigned to reward values and to conditions of informational and rewarding feedback according to the schedule shown in Table 2. For

convenience, the stimuli here and el ewhere are denoted by small letters. In this notation, stimuli i-p were all assigned the reward value 0 and served as the incorrect members of the eight training pairs. Four of the stimuli, e-h, belonged to the Uniform condition and were assigned the reward values of 1, 2, 3, or 4 units. On the occurrence of Uniform items, the associated reward value was always displayed by the experimenter at the end of the trial and was the amount received by the subject as his reward. The items of the critical, "Never-right" condition, denoted c and d in the table, were assigned the reward values 1, 2, or 3, 4 respectively. When the item c-k for example, occurred, if the subject chose card c and guessed any reward value other than 1 or 2, the experimenter displayed 1 or 2 at random; if the subject chose card c and guessed value 1, the experimenter displayed 2, and if the subject guessed 2 the experimenter displayed 1, in either case the subject receiving no reward on the trial. Conditions were analogous for card d with reward values 3 and 4. Cards a and b belonged to the Random condition, for which one or the other of the assigned reward values was displayed whenever the given card was chosen, regardless of the subject's guess, so that on the average once the subject had learned the values assigned to the stimulus he received reward on half the trials.

Training procedure

At the beginning of the training series the subject was instructed as to how trials would be run, that his task was to attempt to choose the correct card of each pair and to guess the associated reward value, and that he would receive the reward displayed if he chose a correct card and had guessed the correct reward value. His objective was to

Table 2

Design of "Never-Right" Experiment

Stimulus	Value Assigned	Informational Condition	Reward on Correct
a	1,2	Random	1,2
b	3,4	Random	3,4
c	1,2	Never Right	0
d	3,4	Never Right	0
e	1	Uniform	1
f	2	Uniform	2
g	3	Uniform	3
h	4	Uniform	4
i-p	0	Uniform	0

attain as large a total of reward points as possible during the experiment. It was explained that the reward he received on each correct trial would be entered in his account by means of the electric counters and that he would know when he received a reward because he could hear the counter operate. On each trial when the subject did in fact choose the correct card and guess the correct reward value, the counter was operated with a clearly audible buzz, though actually no point totals were reported to the subject. The trials were subject-paced, with exposures of stimuli lasting until the subject responded, and with no intervals between trials or cycles except the time required to change cards. On any one trial the two cards of an item were presented, the left-right order being determined randomly; the subject chose a card and guessed a reward value, the experimenter displayed the reward value called for by the procedure on that trial; if the subject was correct in both his choice and his guess the counter was operated to indicate that the subject received the reward displayed. Training continued until the subject went through an entire cycle of 16 items (each of the 8 items appearing in each of the two left-right arrangements), with correct choices and correct guesses of reward values on all trials of the cycle, the correct guess for a Random or a Never-right item being defined as either of the reward values assigned to that stimulus.

Test procedure

After the training criterion had been met, it was explained that the subjects would now be given a series of trials in which the cards involved in the preceding series would be recombined in various ways

but without any other change except that the subject would no longer be shown the reward values at the end of the trial. The counter was operated on all test trials. The subject was given to understand that reward contingencies would be the same as before but that the information was being withheld for purposes of the test. The test series comprised two cycles; during the first cycle all 8 of the correct stimuli of the training series occurred in all possible pairs with one left-right orientation and during the second cycle all of these pairs occurred with the other left-right orientation. Thus there was a total of 80 observations for each test.

Results

Training data

To be credited with a correct response on any trial during the training series, the subject had to choose the correct card of the pair and to guess the correct reward value, in the case of Uniform items, or a value belonging to the assigned pair for the Random or Never-right items. Mean errors to criterion, with errors scored according to this definition, were computed for each item type. It will be recalled that according to the theory under test the speed of learning should not be significantly related to the reward value, or mean reward value, assigned to the correct alternative of an item.

Within the set of Uniform items, no systematic relationship between mean errors and reward value was observed, the means being 9.45, 7.40, 9.05, and 5.35, for items with correct reward values of 1, 2, 3, and 4 respectively. Similarly, errors were not related systematically

to mean reward value within either the Random or the Never-right condition; in the former instances mean errors were 5.62 and 6.70 for the 1, 2 and 3, 4 items and in the latter 6.92 and 5.88 for the 1, 2 and 3, 4 items, respectively. No direct comparison can be made between the Uniform items and the others since the likelihood of guessing the reward values within the assigned subset would be different in the two cases. It is of special interest to note, finally, that there was no appreciable difference in rate of learning, as measured by errors to criterion, between the Random and Never-right conditions, indicating that the rate of learning was a function primarily of the information given regarding associations between stimuli and reward values rather than by reward actually received for correct responses.

Test data

The principal results of the test series are summarized in Table 3 in terms of the proportion of times that a card belonging to any given condition was chosen over cards belonging to each of the other conditions. The values are presented in terms of the proportion of choices of the row entry over the column entry for each cell, with the marginal proportion of times that each type of card was chosen over all competitors indicated in the column of mean proportions at the right.

Considering first the marginal proportions, for the Uniform items the marginals line up monotonically and almost linearly with reward value. Further, the marginals for the Random and Never-right conditions fall between the corresponding pairs of the Uniform conditions, (for example the marginal mean for Random 1, 2 falling midway between the

values for Uniform 1 and Uniform 2) and in fact very close to the values that would be predicted by interpolation if the 1, 2 items were assigned values of 1.5 and the 3, 4 items values of 3.5. Comparing the Random and matched Never-right conditions, there is virtually no difference for the 3, 4 items and a small difference in favor of the Random condition for the 1, 2 items (probably reliable in view of the quantity of data, though no simple statistical test is available).

Considering the interior of the table, within the Uniform condition accuracy of the pair-wise choices was extremely high, indicating that the learning criterion was effective and that there was negligible retention loss between the training and test series. It will be noted that virtually all of the few errors occurring within the Uniform condition involved items with adjacent reward values.

Proportions of choices of a given Random item over the different Uniform items yield steep and monotonic functions, with the 50% point crossed in each instance between the two Uniform values corresponding to those of the given Random item and with a mean difference of about 70 percentage points across these two items. Choices of Random 3, 4 over Random 1, 2 are comparable to those for Uniform 3 over 1 or 4 over 2.

The Never-right items behaved precisely like the Random items in all of these respects, yielding a similar function versus the Uniform items, with crossovers at the same places and an average of about 72 percentage points difference at the crossover. Similarly choices of Never-right 3,4 over Never-right 1,2 (.99) are virtually identical to those of Random 3,4 over Random 1,2, Uniform 3 over Uniform

Table 3

Choice Proportions on Test Series of
"Never-Right" Experiment

Stimulus	Condition		a	b	c	d	e	f	g	h	av.
a	R	1,2		.02	.61	.05	.85	.19	0	0	.25
b	R	3,4	.98		1.00	.44	1.00	.95	.79	.05	.74
c	NR	1,2	.39	0		.01	.76	.04	0	0	.17
d	NR	3,4	.95	.56	.99		.95	.92	.74	.01	.73
e	U	1	.15	0	.24	.05		.04	0	0	.07
f	U	2	.81	.05	.96	.08	.96		0	.01	.41
g	U	3	1.00	.21	1.00	.26	1.00	1.00		.02	.64
h	U	4	1.00	.95	1.00	.99	1.00	.99	.98		.98

1 and Uniform 4 over Uniform 2.

Finally, with regard to a direct comparison of Random versus Never-right items with matched reward pairs, the mean preference was only .52 for the former over the latter, the Random condition having a small advantage at 1,2 and the Never-right condition at 3,4.

Conclusions

It seems clear, from the tests involving pairs of Uniform items especially, that the reward values used were extremely effective in producing significant and systematic effects upon performance. Thus the detailed similarity in patterns of test results for the Random and Never-right conditions and the small differences between the two, wherever matched comparisons were possible, would seem to be of some import regarding one's interpretation of the manner of action of rewarding aftereffects. For items in the Never-right condition, the subjects had opportunity to learn relationships between stimuli and sets of reward values. Clearly this learning occurred and is sufficient to account for the performance on test combinations involving stimuli from Never-right items. On these items the subjects had never received rewards for correct responses during training, and thus there was no opportunity for direct strengthening of associative connections in the manner assumed in the traditional law of effect. Evidently the lack of such opportunity was no appreciable handicap to either learning or test performance. While a strengthening effect cannot be ruled out, the absence of positive support for it in our data is rather striking.

Interpretations of Standard Learning Experiments

Reviewing some of the well established findings of the standard types of human learning experiments may be useful, not so much for the purpose of testing the present theory, which is better accomplished by specially designed new experiments, but for elucidating the intended meaning and mode of application of the concepts more adequately than can be done in a brief summary statement of assumptions.

Thorndikian experiments

The interpretation of Thorndike's lengthy series of trial-and-error learning experiments involving various manipulations with rewards and punishments is a particularly natural first exercise since Thorndike, in effect, began the task himself, discussing a number of his results in terms of what he called the "representational" theory. This theory, which Thorndike believed he was able to refute on the basis of his data, conforms quite closely to the present formulation up to a point. By a "representational" or "ideational" conception, Thorndike referred to the possibility that an image or similar central representation of the rewarding or punishing event might become associated with the stimulus and response terms of an item occurring in an experimental situation. Then, upon later trials, presentation of the stimulus would call up representations of the response and aftereffect previously associated with it, thus putting the learner in the position to choose the response leading to the more desirable outcome. Of the three sources of evidence that Thorndike thought to weigh against the representational theory, two are largely matters of conjecture and even down to the

present can scarcely be evaluated except as a matter of opinion based on intuition or common sense observation. The first of these was the fact that questioning of his subjects did not evoke many verbal reports to the effect that they experienced images of the outcomes of previous trials during the course of his experiments. In view of the notorious fallibility of an individual's verbal accounts of his experiences, this source of evidence needs to be supplemented by experimental tests of learners' ability to recall rewarding and punishing outcomes in order to be taken at all seriously. This task Thorndike seems not to have undertaken seriously.

Results of a recent study conducted in my laboratory are of some interest in this regard, however (Estes, 1966). Following a series of learning trials in which rewards, calibrated in terms of point values, were given for differential responses to a number of stimulus combinations, and a subsequent series of transfer tests, subjects were presented with the stimuli singly and asked to report the previously associated reward values from memory. From the data of this final test it was apparent, firstly, that the subjects could recall the reward values with considerable accuracy, but also that there was a gradient of uncertainty, much like the generalization gradient associated with conditioned stimuli. Thus it is clear that subjects can retain considerable information concerning associations between stimuli and rewards without necessarily manifesting the veridical recall that might be expected if presentation of stimuli on the recall tests evoked images of previously experienced outcomes.

Thorndike's second objection to representational theory seems scarcely open to experimental attack at all. He identified the recall, or reinstatement, of an associative connection between previously experienced events with the occurrence of an image, as the term is commonly understood, then argued that in everyday life situations, such as the performance of skilled athletes, responses follow one another too rapidly to be mediated by images.

The third point involves the empirical generalization established by numerous of Thorndike's researches that an announcement of "right" is a more effective modifier of response repetition than an announcement of "wrong," together with the argument that according to a representational theory there should be no difference since announcements of "right" and "wrong" should be equally likely to be recalled on later occasions. The empirical facts on this matter seem beyond serious question but the inference drawn from them is not. The weakness in the latter has been clearly exposed by Buchwald (1967), who has analyzed the use of "right," "wrong," and nothing (i.e., neither "right" nor "wrong") as reinforcers following verbal responses in terms very close to those of the theory presented in this paper. If a learner's performance on a given trial in this type of experiment is determined by information concerning stimulus-response-outcome relationships of previous trials involving the same item, then it can readily be seen that two of these factors work together to produce repetitions on Right items but tend to cancel each other on Wrong items.

When a subject makes a given response to a stimulus and this event is followed by "right," he has opportunities on that trial to

learn the correct stimulus-response relationship and also the relation between these and the outcome, both of which contribute to the repetition of the response upon later occurrences of the stimulus. On a trial when the subject makes a given response and it is followed by "wrong," he has opportunity to learn the stimulus-response relation, which other things equal increases the likelihood of making that response to the stimulus, and the relation of these to the outcome "wrong," which would lead him to avoid repetition. In the former case neither of the learned relationships would oppose repetition of the response that had been followed by "right;" whereas in the latter case recall of the response without the outcome would tend to lead to increased repetition and only simultaneous recall of the previous response and outcome would lead to an avoidance of repetition. This analysis appears to meet Thorndike's most substantial objection to the representational theory, and we shall see in a later section that the analysis receives independent support from experiments involving novel manipulations of stimulus-response-outcome relationships.

An experimental relation which emerged as a major variable in Thorndikian experiments, and which kept Thorndike himself and later law of effect theorists continually skating on the edge of self-contradiction, is that of "belongingness." As used by Thorndike (1931, 1935) this term referred to two different empirical phenomena. The first was that if during the learning of a task, such as a list of paired-associate items, the learner was suddenly given a monetary reward with some such comment as, "This is a bonus for faithful work," but

with no indication that the reward was related to the response of the preceding trial, the reward would have little or no effect upon the probability that the preceding response would be repeated upon the next occurrence of that item.

The other manifestation of belongingness has nothing special to do with reward or punishment but rather concerns the conditions used to test for learning following presentation of a given set of material. Suppose, for example, that a subject is presented with a series of pairs of letters and digits, then at the end is tested for recall. He will be much more likely to be able to give the correct digit when presented with a letter from the preceding list than he will be to give the letter belonging to the pair which occurred at position N if presented with the digit member of the pair that occurred at the position N-1.

Thorndike proposed to account for these findings in terms of a principle of belongingness, according to which rewarding aftereffects are more influential in strengthening preceding stimulus response associations if both the stimuli and responses, and responses and rewards are related in a way that is meaningful to the subject. His principle seems most foreign in spirit to a theory which depends heavily for its support upon the contention that rewards exert their effects automatically upon preceding stimulus-response associations independently of the learner's awareness of the contingencies involved. Within the present theory, both aspects of the phenomena subsumed under belongingness are special cases of the general principle that later recall of a stimulus-response relationship is facilitated by any conditions leading the learner to perceive the constituent stimulus-response-outcome

relationships contiguously and to rehearse these during the interval between training and testing.

Another principal source of support for the classical interpretation of Thorndikian experiments was the phenomenon of "spread of effect;" that is, the observation that occurrence of a reward following a particular response pair in a series of items frequently led to increased probability of repetition of the responses belonging to neighboring items in the series. The theoretical notion was that the strengthening effect of the reward spread from the rewarded connection to others that were active contiguously in time whether before or after the occurrence of the reward. However, since the concept of spread of effect was an independent assumption developed expressly to account for the phenomenon in question, it scarcely seems logical that the occurrence of the phenomenon can be taken as independent evidence for the law of effect.

The somewhat tortuous literature concerning the spread of effect (see, e.g., reviews by Postman and Sassenrath, 1958; Postman, 1962) leaves one with the net impression that after various artifacts in the earlier studies have been eliminated there remains a genuine effect, though not as large or pervasive a one as assumed by Thorndike. Within the present formulation the observed phenomenon of "spread" has nothing to do with the action of aftereffects, but rather is to be interpreted as a manifestation of stimulus generalization. When the reward occurs at a given point in a list of items, there is opportunity for the learning of relationships between the stimulus member of the item and the reward and also between contextual cues associated with the given position in the list and the reward. The result of this learning will

generalize to temporally contiguous items which share contextual cues with the rewarded one. According to the present theory, learning of the neighboring items will be unaffected, but performance will be modified on a subsequent trial when the common contextual cues have increased probabilities of evoking anticipation of reward.

Operant conditioning of verbal behavior

The extensive literature concerning modifications of verbal behavior by techniques analogous to those of operant conditioning has gone through two principal phases. In the first, it was initially demonstrated by Greenspoon (1955) that rate of occurrence of verbal responses could be modified if, during the flow of spontaneous speech, words having certain common properties were followed by events which might be expected to have rewarding properties, for example the utterance of "good" by an experimenter. Following this demonstration a large number of studies sought to show that procedural manipulations, for example use of fixed ratio or fixed interval reinforcement schedules, discriminative relationships between stimuli and reinforcements, and the like, would produce effects on human verbal behavior similar to those already well documented with respect to the control of bar pressing behavior in rats and key pecking in pigeons by scheduling of food rewards. A measure of success attended these efforts, and reviews of the literature by Krasner (1958), and Salzinger (1959) provided some basis for the conclusion that the processes involved might be basically similar in the human and animal cases.

The assumption that the reinforcements in these experiments exerted their effects via a direct strengthening of response tendencies,

as assumed in Skinner's(1938)principles of operant conditioning, and in Thorndike's law of effect, depended primarily upon the observed similarities in some phenomena between human and animal experiments, and, perhaps more importantly, upon the fact that in some earlier studies questioning of subjects failed to yield evidence that they were aware of the response-reinforcement contingencies. As in the Thorndikian situation, however, later studies (e.g., Spielberger,et al., 1966)approached the question of awareness with more exacting methods and yielded rather clear evidence of strong correlations between indications of awareness of reinforcement relationships and effects of these reinforcements upon response tendencies.

Within the present theory, the operant conditioning experiments are to be interpreted in much the same way as the Thorndikian variety. Rewarding aftereffects should be expected to produce changes in rate of occurrence of verbal utterances in a free responding situation only to the extent that associations between the verbal utterances and the rewarding consequences are learned by the subjects. Although awareness of the constituent relationships on the part of the subjects would not be a necessary condition for this learning, it would be expected to be a correlated variable since the conditions which give rise to awareness are in general the same as those which tend to lead subjects to attend to stimulus-outcome relationships and thus to be in a position to learn them. In the free responding situation, even though the rewarding event may follow a certain verbal response closely in time, since the activity continues without interruption, conditions are not favorable for the subject to perceive the response-reinforcement relation and even less favorable for rehearsal of this association. Thus, it is not surprising

that studies using the free responding procedure have generally found the effects of rewards to be slight and unreliable.

By contrast, in a procedure introduced by Taffel (1955) the experiment is divided into discrete trials with a stimulus presented at the beginning of each trial; the subject chooses a response from a specified set of alternatives, and the rewarding event follows the response which is scheduled for reinforcement. Changes in relative frequency of responding as a function of reward are much larger and more reproducible than in the free responding situation. The change in conditions from the less to the more effective procedure, it will be noted, involves little if any change in the temporal relation of response to reward, but major changes in the likelihood that the subjects will perceive the constituents of the stimulus-reward relationship and in the opportunity for rehearsal of this association between trials.

Magnitude of reward

Within the present theory rewards enter into learning simply as associative elements, or, in other terms, as items of information. After such learning has occurred, associations between stimuli, responses and rewards are an important determiner of performance, as discussed in previous sections of this paper. It should be noted, however, that we have been concerned in this discussion only with specific effects of reward. Although it is scarcely feasible at this stage to incorporate them into a formal theory, there are well known nonspecific effects of rewards which must be taken into account in designing and interpreting experiments concerned with this variable.

In particular, when comparisons are made between different subjects or groups of subjects who learn under different reward conditions, one must recognize the possibility that rewards, like any other motivational condition, may influence subjects' tendencies to attend to material presented and to rehearse stimulus response relationships. Thus, when Thorndike (1935) gave his subjects periodic sums of money which were not associated with particular stimulus response associations, there was no differential strengthening of the responses which had happened to precede these unexpected rewards, but in some instances there was a modification of behavior over the experiment as a whole in comparison to other experiments which did not involve monetary rewards. A better controlled comparison was provided in a study by Thorndike and Forlano (cited in Thorndike, 1935) in which different groups of subjects learned sets of trial and error items with, for one group, monetary reward being associated with correct responses versus none for incorrect responses. Some increase in rate of learning was observed for the group receiving monetary rewards. However, when similar comparisons were made between different items for the same subject, the outcome was quite different. In a study by Rock (cited in Thorndike, 1935), subjects exhibited no differences in learning rates among items of a set of trial-and-error problems when the amount of monetary reward for correct responses, in each case paired with zero reward for incorrect responses, varied over items for each subject.

This negative result led Thorndike to the postulation of a threshold of what he termed the confirming reaction of a reward. According to this notion any reward sufficient to exceed the threshold

would produce a facilitating effect upon learning but further variations in reward value above the threshold would have no differential effects. Although this ad hoc notion could accommodate the results of Rock's study, it would be quite out of line with data of studies such as that of Keller, et al. (1965). In the latter, the rates of approach of learning curves to their asymptotes were found to depend strongly upon reward magnitude when different items in a set had different combinations of reward for the correct and incorrect responses (this result, as noted above, holding only under a noncorrection procedure).

Taking the various types of experiments all into account, one is led to the view, not that each involves some threshold, but rather that effects of reward magnitude are qualitatively different for different combinations of dependent variables and experimental arrangements. As predicted by the present theory, variations in reward magnitude, entailing variation in facilitative feedback, quite uniformly generate corresponding variation in response speeds, or reaction times. However this appears to be strictly an effect on performance, for, as seen in the data of Keller, et al. cited above, the relationship is as pronounced at asymptote as at earlier stages of learning. In contrast, variations in reward magnitude affect relative frequency of correct responding only under conditions such that different amounts of reward convey different amounts of information regarding correctness of response. The accumulating support for this last generalization is by no means limited to my own series of studies. One recent example arising in a quite different context is a study of verbal conditioning by Farley and Hokanson (1966). Rewards varying in monetary value from zero to one cent were assigned

orthogonally, by instructions, different information values relative to correctness of response. Steepness and terminal levels of acquisition curves were directly related to the latter variable but independent of the former.

Delay of reward

The interval between the occurrence of a response and the following reward has been one of the most conspicuous experimental variables in the study of reinforcement. In view of the very extensive literature upon this variable in animal learning and conditioning, it was only natural that attention should turn at an early stage to attempts to reproduce some of the classical animal findings with human subjects. Further, since the classical law of effect is defined in terms of proximity of response and reward, one of the principal consequences of that conception is a gradation in learning rate as a function of delay of reward.

Within the present theory, the interval between the response per se and subsequent rewards is immaterial, although delay of reward may under some circumstances exert effects on rate of learning by modifying the opportunity for learning of associations between the reward and either the stimulus which evoked a given response or stimulus characteristics of the response iteslf. In experiments conducted according to the classical "simultaneous" paradigm, that is, with two or more stimuli presented on a trial and the subject required to select one of these, the interval between response and reward may modify learning rate indirectly if activities occurring during the delay interval lead to the learning of interfering associations. In experiments conducted

according to the "successive" paradigm in which, for example, response 1 is correct for stimulus A and response 2 for stimulus B, with only one stimulus or the other occurring on any one trial, it is the stimulus compound including elements of the presented stimulus, A or B, and the correct response to it which must enter into a learned association with the rewarding event. Thus, in these experiments, and also in experiments on motor skill and the like where the task is primarily one of response selection to a constant stimulus, the degree of discriminability between the stimulus aspects of the alternative responses will be an important variable.

To the extent that the subject can hold representations of the relevant stimuli and responses in immediate memory during the interval of delay, the length of delay interval before occurrence of the reward should be immaterial. Thus we are prepared for the ineffectiveness of delay generally found for simple motor learning tasks with adult human subjects (Bilodeau and Bilodeau, 1958). More importantly, the present interpretation fits well with the finding of Saltzman (1951) that, with the same stimuli, responses, and intervals, delay of reward had no effect under conditions which made it easy for subjects' immediate memory to span the delay interval but produced a significant decremental effect when the load on immediate memory was increased.

In the recent very extensive and carefully controlled study of paired-associate learning in adults by Kintsch and McCoy (1964), conditions were arranged so as to permit delays of 0, 4, or 8 sec. while insuring adequate opportunity for the subject to associate the stimulus and response members. On each trial the subject was presented

with a nonsense syllable stimulus, made a choice of a right or left response key, and then after an interval observed a reinforcing light above the correct key. In Experiment 1, the stimulus was repeated at the time of the feedback light; in Experiment 2 the stimulus was not repeated but was originally presented for a sufficiently long interval to provide the subjects adequate opportunity to encode the stimulus in immediate memory and maintain it over the delay interval. In both cases errors to criterion of learning decreased slightly as a function of the delay of reinforcement.

In studies conducted with children, delay of reward has usually been a more effective variable, though with the conditions of substantial effects generally pointing to the importance of stimulus rather than response-reward relationships. For example, in a study by Brackbill, et al. (1962) children were shown a picture at the beginning of each trial, made a choice of key pressing responses, then were presented with a light over the correct key after a delay interval. However, no measures were taken to insure that the children observed the picture at the end of the delay interval when the reinforcing light appeared, which may well account for faster learning with shorter delays. Even so, retention of the learned relationships was poorer after shorter delays, which is scarcely in line with any theory requiring greater strengthening of associations with shorter delay of reward. And again, in a study by Hockman (1961), children were presented with a stimulus light on each trial and had to select the response button to turn off the light, then receive information as to correctness or incorrectness

after a 0, 10-sec., or 30-sec., delay interval. When there were only two alternative stimuli, orange versus green lights, corresponding to the two response alternatives, there was essentially no effect of delay upon rate of learning. When the number of stimuli was increased to 3, a red or an orange light corresponding to one response and a green to the other, but with no change in the response alternatives, learning was somewhat retarded by the 10- and 30-sec. delays. As implied by the present theory, the effects of the same delays upon the same responses differed in accordance with the discriminability of the stimuli, and therefore with the difficulty of maintaining distinctive representations of the stimuli in memory over the delay intervals.

Perhaps of more diagnostic value with respect to alternative formulations than any of the standard experiments is a recent study by Buchwald (1967), designed especially to demonstrate clearly that delay of reward or of punishment produces effects simply according as the delay entails changes in correlated variables which modify the probability of associative learning. In particular, Buchwald deduced from his analysis of the Thorndikian situation, in which the rewarding or punishing events are indications of right or wrong following the subject's response, that it should be possible to contrive a situation in which delay of reinforcement would actually facilitate learning.

College student subjects were given two trials on each of several lists of common English words with instructions to attempt to learn the responses which went with them. The subject was required to choose a digit as his response each time the stimulus was presented. Under an

Immediate condition, the experimenter said "right" immediately after the subject's response on a Right item and "wrong" immediately after the subject's response on a Wrong item; then on trial 2 the subject attempted to give the correct response immediately upon presentation of the stimulus member of each of these items. Under the Delay condition, the experimenter said nothing following the subject's response on trial 1 but said "right" immediately after presenting the stimulus for a Right item on trial 2; and said "wrong" immediately after presenting the stimulus for a Wrong item on trial 2.

From the standpoint of law of effect theory, the relevant variable is delay of reward and punishment, with the Immediate condition being favorable for learning and the Delay condition involving a very large delay of reward or punishment by ordinary standards and thus being presumably exceedingly unfavorable. According to the present theory, which in application to this situation corresponds in essentials to Buchwald's interpretation, the subject's performance on trial 2 is a function of his ability to recall relationships between the stimulus member of the item and the response which he made to it on trial 1 and between the stimulus member and the rewarding or punishing outcome, if any, given by the experimenter. In the Immediate condition, the subject's response and the experimenter's indication of right or wrong occur immediately following the stimulus on trial 1 and the subject will be in a position to repeat the response if it is correct and suppress it if it is wrong on trial 2 if he remembers both of the constituent relationships. Under the Delay condition, the subject again must recall

his previous response to a given stimulus when it occurs on trial 2 in order to behave appropriately, but the experimenter's indication of right or wrong is supplied at the beginning of trial 2, thus eliminating any necessity for the subject to have learned and recalled this constituent relationship. The principal results were that repetition of a trial 1 response was increased slightly for the delay over the immediate right condition and was reduced substantially more by the delayed wrong than by the immediate wrong condition. Further, in a follow-up study with some modifications in procedure, both results were replicated, but with a substantially greater advantage for delayed right over immediate right in the second experiment (Buchwald, personal communication 1967).

Summary

A review of the literature on reward in human learning reveals major inadequacies in the law of effect interpretation which become accentuated as experimental designs depart from classical paradigms. Although it is indeed the case that in many everyday life situations, and in many standard experiments, subjects come with experience to select responses which lead to reward over those which do not, the conclusion does not follow that there is a direct and fundamental connection between the occurrence of the reward and the establishment of a learned association between previous stimuli and responses. New types of experiments contrived to bear especially upon this issue indicate, in fact, that the associative learning process is independent of rewarding or punishing aftereffects. Further, the increasingly complex pattern of empirical relationships involving such major variables as delay and magnitude of reward are not satisfactorily interpreted by

any extant theory cast in the law of effect framework.

The alternative theoretical framework outlined in the present paper embodies the following principal assumptions.

1. Associative learning is a function solely of conditioning by contiguity. When any two behavioral events (whether they are termed stimuli or responses is immaterial) occur in succession there is some probability that an association between the two members will form on an all-or-none basis. The result of formation of an associative linkage is that on a later occasion the occurrence of the first event calls into memory a representation of the second.

2. Recognition performance depends solely upon this learning by contiguity.

3. If the second member of an association is a response which has occurred on the previous occasion, the overt occurrence of this response upon a later test requires not only activation of the association but also facilitative feedback, which arises as a function of anticipation of reward.

4. Modification of learned performance by rewards involves two stages. Firstly, an association must be established between the stimulus member of a stimulus-response event and the subsequent reward. Secondly, upon the subsequent occurrence of the stimulus member, the representation of the reward must be brought into memory by activation of the learned association. This anticipation of reward generates facilitatory feedback. If at the same time the previous response is recalled, the effect of the feedback is to cause the recalled response to be made overtly.

5. The immediate effect of facilitatory feedback following presentation of a stimulus is to increase the probability that the associated response will be made overtly during any given interval of time, and thus on the average to reduce the reaction time.

6. In a choice situation, in which the subject must select the stimulus to which to respond from a set of two or more, the subject is assumed to scan the available stimuli and, in general, to respond overtly to the one with which the highest reward has been associated during previous learning experiences. Thus, in selective learning situations, the function of rewards may be characterized as one of stimulus amplification. Different stimuli become associated with different rewards and the consequence is that the stimuli then carry differential weights in the determination of response, either in the same or in new situations.

This theoretical schema has received relatively direct support from an experiment designed to permit the learning of stimulus-reward associations in the absence of any opportunity for the direct strengthening of stimulus-response associations by the rewards. Further, the consideration of several major areas of research on human learning indicates that effects of major variables such as delay and magnitude of reward, as well as information value (relative to correct versus incorrect responding), can be organized and interpreted within this framework. This is not to say, however, that the theory is complete and satisfactory in all respects. In fact, it is definitely limited as presently formulated in that the task remains to develop an adequate conceptualization of motivational conditions in human learning and their relationships to rewards and punishments.

References

Battig, W. F., & Brackett, H. R. Comparison of anticipation and recall methods in paired-associate learning. Psychol. Reports, 1961, 9, 59-65.

Bilodeau, E. A., & Bilodeau, I. McD. Variation of temporal intervals among critical events in five studies of knowledge of results. J. exp. Psychol., 1958, 55, 603-612.

Bower, G. H. A model for response and training variables in paired-associate learning. Psychol. Rev., 1962, 69, 34-53.

Brackbill, Y., Bravos, A., & Starr, R. H. Delay improved retention of a difficult task. J. comp. physiol. Psychol., 1962, 55, 947-952.

Buchwald, A. M. Effects of immediate vs. delayed outcomes in associative learning. J. verb. Learn. verb. Behav., 1967, 6, 317-320.

Dulany, D. E., Jr. The place of hypotheses and intentions: An analysis of verbal control in verbal conditioning. In C. W. Eriksen (Ed.), Behavior and Awareness. Durham: Duke Univer. Press, 1962, 102-129.

Estes, W. K. The statistical approach to learning theory. In S. Koch (Ed.), Psychology: A study of a science. Vol. 2 New York: McGraw-Hill, 1959, 380-491.

Estes, W. K. Transfer of verbal discriminations based on differential reward magnitudes. J. exp. Psychol., 1966, 72, 276-283.

Estes, W. K. Outline of a theory of punishment. Tech. Rep. No. 123, Stanford Univer., 1967. (To appear in Punishment and Aversive Behavior, R. M. Church and B. A. Campbell (Eds.), Appleton-Century-Crofts, in press.

Estes, W. K., & Da Polito, F. Independent variation of information storage and retrieval processes in paired-associate learning. J. exp. Psychol., 1967, 75, 18-26.

Farley, J. A., & Hokanson, J. E. The effect of informational set on acquisition in verbal conditioning. J. verb. Learn. verb. Behav., 1966, 5, 14-17.

Greenspoon, J. The reinforcing effect of two spoken sounds on the frequency of two responses. Amer. J. Psychol., 1955, 68, 409-416.

Humphreys, M. S., Allen, G. A., & Estes, W. K. Learning of two-choice, differential reward problems with informational constraints on payoff combinations. J. math. Psychol., 1968, in press.

Hockman, C. H., & Lipsitt, L. P. Delay-of-reward gradients in discrimination learning with children for two levels of difficulty. J. comp. physiol. Psychol., 1961, 54, 24-27.

Keller, L., Cole, M., Burke, C. J., & Estes, W. K. Reward and information values of trial outcomes in paired-associate learning. Psychol. Monogr., 1965, 79 (Whole No. 605).

Kintsch, W., & McCoy, D. F. Delay of informative feedback in paired-associate learning. J. exp. Psychol., 1964, 68, 372-375.

Krasner, L. Studies of the conditioning of verbal behavior. Psychol. Bull., 1958, 55, 148-170.

Nunnally, J. C., Duchnowski, A. J., & Parker, R. K. Association of neutral objects with rewards: Effect on verbal evaluation, reward expectancy, and selective attention. J. Pers. soc. Psychol., 1965, 1, 274-278.

Nunnally, J. C., Stevens, D. A., & Hall, G. F. Association of neutral objects with rewards: Effect on verbal evaluation and eye movements. J. exp. child Psychol., 1965, 2, 44-57.

Postman, L. Rewards and punishments in human learning. In L. Postman (Ed.), Psychology in the making. New York: Knopf, 1962.

Postman, L., & Sassenrath, J. The automatic action of verbal rewards and punishments. J. gen. Psychol., 1961, 65, 109-136.

Salzinger, K. Experimental manipulation of verbal behavior: A review. J. gen. Psychol., 1959, 61, 65-94.

Saltzman, I. J. Delay of reward and human verbal learning. J. exp. Psychol., 1951, 41, 437-439.

Skinner, B. F. The Behavior of Organisms. New York: Appleton-Century-Crofts, 1938.

Spielberger, C. D. The role of awareness in verbal conditioning. In C. W. Eriksen (Ed.), Behavior and Awareness. Durham: Duke Univer. Press, 1962, 73-101.

Spielberger, C. D., Bernstein, I. H., & Ratliff, R. G. Information and incentive value of the reinforcing stimulus in verbal conditioning. J. exp. Psychol., 1966, 71, 26-31.

Taffel, C. Anxiety and the conditioning of verbal behavior. J. abnorm. soc. Psychol., 1955, 51, 496-501.

Thorndike, E. L. Human Learning. New York: Appleton-Century-Crofts, 1931.

Thorndike, E. L. The Psychology of Wants, Interests, and Attitudes. New York: Appleton-Century-Crofts, 1935.

Witryol, S. L., Lowden, L. M., & Fagan, J. F. Incentive effects upon attention in children's discrimination learning. J. exp. child Psychol., 1967, 5, 94-108.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA · R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author) Stanford University Institute for Mathematical Studies in the Social Sciences, Stanford, California 94305	2a REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b GROUP

3 REPORT TITLE

Reinforcement in Human Learning

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report, -- December 1967

5. AUTHOR(S) (Last name, first name, initial)

Estes, W. K.

6. REPORT DATE	7a TOTAL NO OF PAGES	7b. NO. OF REFS
12/20/67	50	31

8a. CONTRACT OR GRANT NO.	9a ORIGINATOR'S REPORT NUMBER(S)
Nonr-225(73) b. PROJECT NO. NR 154 218 c. d.	Technical Report No. 125 9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Personnel and Training Branch, Code 458 Washington, D. C. 20360

13. ABSTRACT

A review of the literature on reward in human learning reveals major inadequacies in the law of effect interpretation which become accentuated as experimental designs depart from classical paradigms. New types of experiments indicate that the associative learning process is independent of rewarding or punishing aftereffects. The alternative theoretical framework outlined in the present paper embodies the following principal assumptions:

1. Associative learning is a function solely of conditioning by contiguity.
2. Recognition performance depends solely upon this learning by contiguity.
3. Modification of learned performance by rewards involves two stages. Firstly, an association must be established between the stimulus member of a stimulus-response event and the subsequent reward. Secondly, upon the subsequent occurrence of the stimulus member, the representation of the reward must be brought into memory by activation of the learned association. This anticipation of reward generates facilitatory feedback. If at the same time the previous response is recalled, the effect of the feedback is to cause the recalled response to be made overtly.
4. The immediate effect of facilitatory feedback following presentation of a stimulus is to increase the probability that the associated response will be made overtly during any given interval of time, and thus on the average to reduce the reaction time.
5. In a choice situation, in which the subject must select the stimulus to which to respond from a set of two or more, the subject is assumed to scan the available stimuli and, in general, to respond overtly to the one which the highest reward has been associated.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Reinforcement Reward and Punishment Human Learning						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

_____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

_____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

_____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.

DD FORM 1 JAN 64 1473 (BACK)

UNCLASSIFIED

Security Classification