Acquisition of correct choices and value judgments in binary choice learning with differential rewards*

G. A. ALLEN† and W. K. ESTES
The Rockefeller University, New York, N.Y. 10021

Concurrent training was given on a list of items, each comprising a pair of stimuli with assigned numerical reward values. On each trial, S chose one member of the stimulus pair displayed and gave an estimate (judgment) of its value before receiving the appropriate reward. Learning functions for reward judgments were systematically related to differences between high and low values of stimulus pairs but tended to lag behind functions for probability of making optimal choices.

Our purpose in this study was to seek more direct evidence than heretofore available concerning the basis of learning by adult human Ss in a binary choice situation with differential rewards. In the type of experiment under consideration, the S receives training on a list of items, each item comprising a pair of stimuli between which the S must make a choice on each trial, with a reward value assigned to each stimulus. When training is given with full information, the S is shown at the end of each trial both the reward value he receives for the choice he made and also the one he would have received for the other choice on a given item. In the partial information condition, S is shown only the value corresponding to the choice that he made on each trial.

The trend of recent research has seemed to favor a simple informational or cognitive interpretation of learning in this situation (Estes, 1971). Thus, for example, Keller, Cole, Burke, & Estes (1965) found that rate of learning of individual items was directly related to the difference in magnitudes of the assigned rewards under the partial information procedure but that the difference disappeared when the amount of information conveyed per trial was equated by means of the full information procedure. In a study by Humphreys, Allen, & Estes (1968), conditions were arranged so that, if Ss acquired all of the available information, the reward values they anticipated for various choices in some

instances would differ substantially from the reward values they had actually received on previous training trials for these choices. Their choices proved to be related in a more simple and orderly way to anticipated reward values than to magnitudes of reward actually received.

A number of quantitative aspects of choice learning have been accounted for reasonably well by a model which assumes that rate of forming associations between stimulus alternatives and rewards is independent of reward magnitude. The direct relationships between choices and reward values observed under some circumstances are interpreted as performance effects (Humphreys, Allen, & Estes, 1968). However, we should not be satisfied with an assumption when more direct observation is possible. Thus we have contrived a variation on the binary choice experiment which permits repeated memory probes for S's state of information.

In the modified procedure, the list structure is the same as in earlier studies and choices are made on each trial in the same fashion, but S, following his choice on each trial, makes an estimate of the associated reward value before he receives any information or feedback. Except for the additional response required on each trial, the design and procedure of the present experiment replicate those of Humphreys, Allen, & Estes (1968).

By this means, we hope to clarify at least two points of central theoretical interest: (1) whether rate of acquisition of information depends on reward value; and (2) whether S's state of information concerning reward values at various stages of choice learning is a sufficient basis for the observed probability of optimal choices.

SUBJECTS AND APPARATUS

The 40 Ss, all Stanford students or

students' wives, were paid for participation.

The S sat by himself at a table in a soundproofed air-conditioned room which contained the stimulus display apparatus and response keys. On a table 5.5 ft in front of the S were two stimulus display boxes, one on top of the other, masked by a black screen so that only the relevant portions of the displays could be seen. Alphanumeric characters could be displayed in each box. The upper display had a 2 x 12 in. window, through which the stimuli, pairs of words, were displayed. The lower display had a 2 x 1 in. window, located halfway between the positions of the stimulus words, through which the reward values were displayed. On the table immediately in front of S were two columns of 10 response buttons; 9 of the buttons in each column were labeled with the digits 0-8 and the 10th with either the word "left" or the word "right." The 0 button was closest to S and the "left" or "right" button farthest from S in each column.

A control room, adjoining the S room, contained an IBM summary punch and associated storage and timing equipment. The sequence of stimulus-response-reward trials was generated by a preprogrammed input to the apparatus by IBM cards.

STIMULI AND

REWARD COMBINATIONS

The stimuli were 40 three-letter common nouns arranged in two lists of 10 pairs. The pairs of reward values, as in the study by Humphreys, Allen, & Estes (1968), were 8-0, 3-0, 8-5, 7-5, 5-3, 3-1, 7-6, 6-5, 3-2, and 2-1. Within each list, each word pair was assigned to one of these reward combinations without replacement. Each word in a pair was associated with one of the numbers of the assigned reward combination. Thus there were two lists of 10 word pairs with identical reward structures. For each of the lists, henceforth denoted Lists I and II, two variations were used, differing only in the left-right positions of the words in each pair. The words and reward values were so arranged that half of the items had the word carrying the lower reward on the left side of the display screen and half on the right side; the second variation of each list simply reversed these left-right relationships.

The two lists, each with its two variations, formed the input for a computer program which generated and randomized the IBM cards that controlled the stimulus-reward sequences. Sets of five Ss received the same assignments of word pairs to reward values but different presentation orders. The randomization of item presentations

*This study was initiated while the authors were at Stanford University; Michael S. Humphreys participated in the experimental work. Analyses of the data and preparation of this report were supported by Grants MH16100 and GM16735 from the U.S. Public Health Service.

†Now at Miami University, Oxford, Ohio 45056.

Fig. 1. Proportion of correct value judgments on each of the first eight correct choices. Data are pooled by values of Δ , the reward differential between high and low members of a stimulus pair.

was constrained so that at least seven items intervened between two successive occurrences of any given item. Other details of the randomization and presentation procedure were the same as those used by Humphreys, Allen, & Estes (1968).

PROCEDURE

Each S was presented with the two lists successively, each list being presented 16 times. The average presentation time was about 20 min per list, and there was a 5-min rest between lists.

The instructions given the S indicated that when a word pair appeared on the upper display, the S was to respond by selecting the left or right column of response buttons according to which word (left or right) he believed had the higher reward value and then to press a button in the chosen column. If he felt that he knew the associated reward value, he was to indicate this by pressing the proper numbered button; if not, he was to press the "left" or "right" button, indicating that his choice was a guess. It was made clear to the S that the

amount of money he earned was a function of the total number of points he obtained as rewards. Since the total number of points obtained by an S during a session was not accessible to the E until the IBM cards had been processed, a sample of each S's responses and rewards during Trials 21-50 on List II was recorded by hand (without the S's knowledge) and the payments were calibrated so as to be linearly related to the points obtained on these trials. The amounts of money earned per session per S ranged from \$1.60 to \$2.10, with an average of \$1.80.

The trial procedure was similar to that of the anticipation paradigm in paired-associate learning and conformed to the partial information condition of the study by Keller et al (1965). The sequence of events on a trial was as follows: The word pair appeared on the upper display screen and remained on until after the S pressed a response button. As soon as the S pressed the response button, this button was illuminated. After 1.5 sec, the number of points earned on the

trial appeared on the lower display screen. After 1 sec, the stimulus and reward displays and the illuminated key were all turned off together; after a 3-sec dead interval, the next trial began.

RESULTS

In presenting the results, we shall denote choice of the stimulus of a pair carrying the higher reward value as a correct response and choice of the stimulus with the lower reward value as an error; also, for brevity, we shall refer to selection of the "left" or "right" response button as a *guess* and selection of a numbered button as a *judgment*. Selection of the response button labeled with the number corresponding to the reward value of the stimulus on the same side will be denoted a correct value judgment, regardless of whether the member of the stimulus pair chosen on the given trial was the one carrying the higher value.

The numerous statistics concerning correct responses and errors which have been analyzed in detail in the preceding studies are not of primary interest here. Terminal mean levels of correct choice probability are about .80 for List I and slightly over .90 for List II. The mean total errors over all items of Lists I and II in the present study are 5.39 and 3.36, as compared to values of 5.16 and 3.77, respectively, in the data of Humphreys, Allen, and Estes. This and similar correspondences in other statistics suggest that the process of learning correct choices has not been greatly disturbed by the additional requirement of giving reward value judgments in the present study. As in the previous experiments, total errors are generally related inversely to the difference between the high and low reward values of an item but directly to the absolute reward values of incorrect alternatives.

The course of acquisition of correct value judgments is shown in Fig. 1. These curves represent all protocols having at least eight correct choices, with data pooled for each value of Δ , the difference in reward value between the high and the low alternative of a pair. It will immediately be seen that, even with numbers of opportunities for learning equated, the rate of acquisition of correct associations between stimuli and their reward values is by no means independent of value. Further, though there is an across-the-board increase in learning rate from List I to List II, there is no change in the systematic dependence of learning rate upon reward differential.

Rate of learning to make correct judgments is even more uniformly and strongly related to value of the given

alternative than to reward differential. Information bearing on this relationship is presented in Table 1. Again, utilizing all protocols having at least eight correct choices, we have computed mean numbers of correct value judgments over these trials, with the data pooled for each value of the correct alternative. Thus the first row is based on the 8-0 and 8-5 problems, and so on. Estimated standard errors of the means are based on the within-sample mean squares. It will be observed that the means line up in the order of the reward values, with only the 3-value out of place for List I and with no inversions for List II. The likelihood of this ordering arising by chance is well below .01. Further, although an analysis of variance is not feasible, owing to the difficulties of dealing with varying cell frequencies over sets of related measures, the magnitudes of the differences between the mean judgment scores for high and low values are so large relative to their standard errors as to leave no doubt concerning reliability of the observed trend.

In order to ascertain how changes in S's state of information about rewards are related to changes in choice behavior, we first raise the question as to whether any significant preference for the correct alternative of an item appears before S has learned at least one of the reward values. Thus we have determined for each protocol the numbers of correct choices on trials before the first correct value judgment on either alternative. Pooling these data over Ss and over all items in each list, we find an average proportion of .55 correct choices for List I and .56 for List II, based on 2,081 and 1,375 observations, respectively. Within both lists, there are similar direct relations between proportion of correct choices on these "no information" trials and reward differential for an item: mean proportions are .54, .55, .58, and .69 for reward differentials of 1, 2, 3, and 8, respectively.

Before we can interpret this result, we need to exclude the possibility (pointed out by Douglas Medin) that the higher-than-chance proportions might arise from sequences in which S actually had learned the reward value for the incorrect alternative. Particularly when the value is very low, S might henceforth avoid that stimulus and therefore never reveal his knowledge of its value but have to guess at the value of the correct alternative until he learned it also. Two considerations indicate that instances of this sort are rare. First, the average proportions of correct choices prior to the first correct judgment for items with 5 or 6 as the value of the incorrect alternative are

Table 1
Means and Estimated Standard Errors of Number of Correct Value Judgments Over the First Eight Correct Choices

Value of Correct Alternative	List 1		List 2	
	M	SE _M	M	SE _M
8	4.57	.37	6.01	.23
7	3.60	.40	5.74	.27
6	3.48	.42	5.15	.27
5	3.19	.36	5.05	.31
3	3.69	.45	4.91	.27
2	3.07	.42	4.31	.36

quite close to the overall proportions—.54 and .55 for Lists I and II, respectively. Secondly, we rescored the 8-0 and 3-0 items, utilizing only trials prior to the last error in each protocol. On these trials, S could not reasonably be supposed to have known that the incorrect alternative had value 0, since he later chose it and assigned some other value. This analysis yielded proportions correct of .60 for List I and .68 for List II (as compared to .62 and .64 in the preceding analysis).

Evidently, we may conclude that Ss often begin to improve their choice performance on an item before they manifest any knowledge of the assigned reward values. Is it possible that acquisition of information about reward values continues to lag behind the increases in correct choice probability throughout the learning process? To obtain evidence on this possibility, we wish to confine attention to instances in which Ss have met a reasonable criterion of mastery of the choice problem. Therefore, we turn to analysis of the subset of protocols which meet a criterion of five consecutive correct choices.

Taking the last error in each criterion protocol as a point of reference, we have computed the proportion of correct value judgments on subsequent trials. These proportions for both lists are plotted for the five criterion trials in Fig. 2. Clearly the learning of correct reward value largely occurs following the last error. It should be observed, however, that the extremely low proportions correct on the first trial of the criterion run for both lists are attributable in part to protocols in which there were no correct choices and therefore no opportunities to learn the correct reward value prior to the last error. If we consider only criterion protocols with at least one correct choice before the last error, the proportion of correct value judgments on the first criterion trial is .40 for List I and .65 for List II. Thus, even in these cases, Ss usually begin choosing the correct alternative 100% of the time before they can state its reward value.

It is of interest also to ask whether knowledge of the reward values increases progressively over the trials just before S begins his criterion sequence of correct choices. The answer is no. For the same set of protocols considered in the last analysis, proportions of correct value judgments on the last three correct choice trials before the last error are .22, .37, and .26, respectively, for List I and .45, .43, and .47 for List II. Mean deviations of judgments from true values are 1.5, 1.7, and 1.5 on these trials for List I, with a drop to 1.0 on the first criterion trial; corresponding values for List II are 1.1, 0.9, and 0.5, respectively.

Similar constancy obtains for judgments of value of the incorrect alternative over precriterion trials. In List I, for which there are substantial data available over four trials preceding the last error, the mean proportions of error trials on which value judgments were correct over this sequence are .33, .34, .31, .29, and .30 (the last value being the one obtaining on the trial of the last error). Particular significance attaches to the last two points, for which all reward combinations are represented with at least 10 observations each. Quite possibly, a sharp increment in S's information concerning value of the

Fig. 2. Mean proportions of correct value judgments on the first five trials after the last error on each list.

incorrect alternative occurs immediately following the last error, but, of course, no relevant evidence can be obtained from this study.

Further examination of the individual protocols reveals an error pattern which may be of theoretical significance, namely, a substantial incidence of interchanges of reward judgments between the two alternatives of an item. In the precriterion protocols for List I, there were 87 instances in which S made an error after giving the correct value for the correct alternative; in nearly 40% of these instances, the judgment given on the error trial was the value of the correct alternative.

This tendency to interchange reward values of an item persists throughout the learning. To illustrate, let us consider the final four trials on List II. For trials in this block on which correct choices occurred, constituting 90% of the total, approximately 20% of the incorrect judgments were judgments of the reward value for the incorrect stimulus of the given item. For the trials on which incorrect choices occurred, over 40% of the incorrect value judgments were judgments of the value of the correct stimulus for the given item.

It appears that models previously proposed for this type of learning situation are oversimplified in that they fail to provide for the possibility that associations of reward values with the stimulus compound displayed upon each occurrence of an item and the correct assignment of the value, once learned, to the appropriate one of the two alternative stimuli may involve at least partially independent processes.

DISCUSSION

The two-choice differential reward experiment can be regarded as a generalization of verbal discrimination learning, informative feedback being provided by higher and lower reward values rather than "right" and "wrong." Under the partial information procedure of the present study, the differential reward procedure sets a considerably more difficult task for S, since mastering of a problem requires the combining of information obtained on at least two different trials. Similarly, learning under this procedure poses more difficult problems of interpretation.

The problem for theory is accentuated by the fact that our results confute all of the simplifying assumptions of the model which seemed most promising when we began this study. In particular, we have been led to conclude that: (1) the acquisition of correct judgments is systematically related to reward value; (2) there is significant improvement in

probability of correct choice on an item before the first evidence appears that S has learned either of the reward values; and (3) the reward value of the correct alternative of an item is most often learned after the last error.

It becomes clear that something important is missing from a simple cognitive account of choice learning. Before we probed the memory states of Ss directly, it seemed intuitively plausible that adult Ss simply learn stimulus-reward relationships in a paired-associate fashion and base their choices on what they have learned. Further, a model embodying these assumptions provided promising quantitative accounts of choice data (Humphreys, Allen, & Estes, 1968). However, a seed of doubt concerning the adequacy of the model was planted by Medin's (1969) finding that the fits of the model to data could be improved by assuming some improvement in choice behavior prior to acquisition of reward information. This idea has been borne out rather strongly by the results of the present study.

What could be the basis for improvement in choice performance prior to the acquisition of verbalizable reward information? A hint as to the answer may be found in the relationship between rate of acquisition of judgments and reward differential shown in Fig. 1. The basis of this relationship is probably additional rehearsal of the items with higher reward values for the correct choice or with higher reward differentials. Rehearsal of stimulus words together with their values would, of course, tend to increase recall. But even in instances when the rehearsal failed to establish long-term memory for the stimulus-reward association, it would still be expected to increase stimulus familiarity. Further, Medin (1972) reports that Ss often rehearse only the name of a high-valued stimulus rather than the stimulus together with its reward. Just as in verbal discrimination learning (Eckstrand, Wallace, & Underwood, 1966), increased familiarity of a stimulus would be expected to lead to an increase in its likelihood of being chosen on a subsequent trial.

Since S knows the range of possible reward values at the beginning of the experiment, this differential rehearsal could arise on the very first cycle through the list. The differences in acquisition rate reflected in Fig. 1 suggest that when the stimulus chosen yields a reward falling in the upper half of the range, the stimulus is immediately rehearsed, in an amount directly related to its value. When the stimulus yields a reward in the lower half of the range, the other member of

the pair is similarly rehearsed, the amount being inversely related to the observed value.

This hypothesis predicts the very high proportions correct for the 8-0 item and the higher than average proportion correct for the 3-0 item prior to the first correct value judgment. Also, it accounts for the frequent observation (Humphreys, Allen, & Estes, 1968; Keller et al, 1965) that items with high values for the incorrect alternative (for example, 8-5 or 7-6) yield a high incidence of cases in which Ss "absorb" on the incorrect alternative and never meet the criterion of learning.

The generation of differences in stimulus familiarity by differential rehearsal could provide a sufficient basis for the learning of items which combine a high and a low reward (greater and less than 4, respectively, in this study) and for all items including the maximum or minimum values (8 or 0). But something more is needed to account for the fact that intermediate high pairs (e.g., 7-5) are usually and low pairs (e.g., 3-2) virtually always learned to criterion, and at rates not very different from high-low pairs.

The most obvious suggestion is that mastery of items involving high or low reward pairs depends on Ss learning the reward values. This idea receives some support from our finding that Ss do acquire substantial information regarding specific reward contingencies. However, we should note that the requirement of making explicit value judgments in this study undoubtedly led to more learning of reward values than would occur under the usual conditions of choice experiments. And even under the especially favorable circumstances of our study, Ss did not generally exhibit knowledge of both reward values of an item by the end of the learning series.

How can an intermediate high or low pair be mastered if only one of the reward values is learned? The rehearsal-familiarity mechanism may again provide the answer. It is necessary only for S to learn either value of a pair, then on the next occurrence of the item to choose the other member and observe the associated reward, and finally to rehearse the stimulus which has proved to yield the higher reward.

With regard to theories of choice learning, we have come another full cycle. The model which motivated our introduction of the memory probe procedure has proved inadequate in major respects. We have been led to a revised conception of the learning process which accommodates our new findings and which evidently provides a sufficient basis for mastery of all

item types used in these studies. But it remains to be determined whether this process is the one Ss actually use.

Once again, new variations in procedure are needed. First of all, we need to trace the acquisition of information regarding both the correct and the incorrect alternatives throughout learning, preferably in a way which will not perturb the normal course of choice learning. Also, we need to find some means other than verbal reports to check on S's state of reward information. Both of these tasks have been essayed in a follow-up

study soon to be reported (Allen, 1971).

REFERENCES

- ALLEN, G. A. Memory probes during two-choice, differential reward problems. PhD dissertation, Stanford University, 1971.
- ECKSTRAND, B. R., WALLACE, W. P., & UNDERWOOD, B. J. A frequency theory of verbal discrimination learning. *Psychological Review*, 1966, 73, 566-578.
- ESTES, W. K. Reward in human learning: Theoretical issues and strategic choice points. In R. Glaser (Ed.), *The nature of reinforcement*. New York: Academic Press, 1971. Pp. 16-36.
- HUMPHREYS, M. S., ALLEN, G. A., & ESTES, W. K. Learning of two-choice, differential reward problems with informational constraints on payoff combinations. *Journal of Mathematical Psychology*, 1968, 5, 260-280.
- KELLER, L., COLE, M., BURKE, C. J., & ESTES, W. K. Reward and information values of trial outcomes in paired-associate learning. *Psychological Monographs*, 1965, 79(12, Whole No. 605).
- MEDIN, D. L. Models for two-choice differential reward learning. Technical Report, Rockefeller University, 1969.
- MEDIN, D. L. Partial information and choice behavior in differential reward magnitude learning. *Psychonomic Science*, 1972, in press.