JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 6, 891–896 (1967)

Changes in the Memory Operating Characteristic during Recognition Learning[1]

WALTER KINTSCH

University of California, Riverside 92502

AND

WARREN J. CARLSON

University of Missouri, Columbia, Missouri

Four Ss learned to recognize 30 lists of paired associates. Confidence judgments were obtained with all responses. The learning data were in agreement with a Markov model which requires a constant error probability in the initial state. The The confidence judgments were used to construct memory-operating characteristics. The MOC based upon all scores of the first test trial is smooth and symmetric and indicates good discrimination. MOCs were also constructed from scores of the first test trial which were followed by an error on some later trial, and from all scores before the last error. These two MOCs overlapped and were in between the MOC based upon all first-trial scores and the chance line. It is concluded that on trials before the last error Ss possess some information about the learning material but that the amount of information does not increase during those trials.

Recently the technique of constructing Receiver-Operating Characteristics has been applied to the study of memory. Egan (1958), Murdock (1965) and Norman and Wickelgren (1965) have noted the formal similarity between the short-term recognition experiment and the psychophysical experiment for which this technique had originally been developed. Egan described a procedure by means of which confidence ratings which are obtained along with yes-no recognition responses may be used to construct Memory-Operating Characteristics (MOCs). The results obtained with this technique have important implications for any theory of recognition memory. Murdock obtained smooth, symmetric MOCs, very much like those postulated by signal-detection theory. The parameter d' was found to vary with serial position in a reasonable fashion. Norman and Wickelgren found similar MOCs under one of their experimental conditions and asymmetrical curves under another.

A new analytic tool such as the MOC provides information about new aspects of recognition learning and permits us to extend existing models of recognition learning, or forces us to reject them. The purpose of the present study is to investigate whether and how MOC curves are compatible with the model of recognition learning of Kintsch (1966). This model is a Markov model with a nonlearning state, a short-term memory state, and a permanent-storage state. It is formally equivalent to a simple all-or-none model as far as the

[1] This study was supported by grant MH11445 from the National Institute of Health. A report based upon the present data was given by the senior author at the XVIIIth International Congress of Psychology, Moscow, 1966.

success-error response sequences are concerned. A number of features of such sequences were shown to be in good agreement with the model. Of course, such a model will, at best, provide an overall description of the structure of recognition learning. New experimental designs as well as new theoretical approaches are needed to fill in the details of the processes which underlie this structure.

The usefulness of the MOC arises because it permits the separation of the effects of response biases from true learning effects, i.e., increases in discrimination between learning items and distractors. Thus answers may be provided to questions such as "How much discrimination is possible after one learning trial," "Is performance better than chance on trials when, according to the Markov model, items are not yet in the learning state," and, particularly interesting, "Does discriminability change over trials as long as items are not in the learning state?" According to the all-or-none interpretation of the model, as long as an item is not learned, Ss can do no better than to guess. In this case, the answers to the last two questions should be "no." According to the short-term memory version of the model, responses before the last error are determined both by the S's response biases and forgetting from an imperfect memory store. Thus one would expect an MOC well above the chance line which does not change on trials before the learning state is entered.

Method

Subjects. Four University of Missouri undergraduates were paid $1.25 per session for 30 sessions. Two men and two women participated; all were without prior experience in verbal learning experiments.

Materials. Fifteen lists were constructed from moderate-association-value (Glaze 33–55%) CVC trigrams and 15 were constructed from low-association-value consonant trigrams (Witmer 17–25%). For the first 12 days the list length was 14 pairs, for days 13–15 the lists were composed of 18 pairs, while for the final 15 days 22 pairs were used. For both CVC and CCC lists intralist similarity was minimal as the lists were constructed with the requirement that each letter occur about equally often. There were two response alternatives for any given day. These were either a pair of one-digit numbers such as 1 vs. 2 or, equally often, a pair of letters such as A vs. E, or K vs. M. The specific response pair varied from day to day. The letters which were used were always picked after the list was constructed and were two of the least frequently occurring for that day. The trigrams were typed in capitals centered upon 3 × 5 in. white cards. The syllable and the designated correct response were typed on one side of the card and the syllable and the incorrect alternative were typed on the back.

Procedure. Each S learned 30 lists, half of which were CVC trigrams and half CCCs. The sequence of CVC and CCC lists was randomly determined with all Ss receiving the lists in the same order, one list per day.

The first presentation of each list was a study trial with all trigram response pairs correct. Thereafter half of the pairs which were presented were correct and half were paired with the incorrect alternative. Cards were presented manually for about 1.5–2 sec. The Ss were instructed to read aloud each card and then to say either yes or no depending upon whether they thought the trigram was paired with its correct response alternative or not. They also were asked to rate their confidence in their judgment on a 5-point scale, saying 5 if they were sure of their response and 1 for a pure guess. Following their responses Ss were informed whether they had been correct or not. A trial consisted of the presentation of all pairs of a list. Between trials *E* rearranged the cards so that a random half of all items was paired with the correct alternative and the rest with the incorrect alternative. To avoid practice by S during the intertrial interval S was asked to count backwards by some number, writing the successive remainders on paper. The average duration of the intertrial interval was about 90 sec. Trials were continued until S reached a criterion of two successive correct trials.

Results

Some of the results are typical of this kind of learning situation and need only be mentioned briefly. Learning rate increased over the 30 sessions for each S. Averaged over Ss, the mean number of errors per item was .73 for the first 15 lists and .45 for the second half of the lists. Trigrams were learned somewhat more slowly than syllables, the mean number of errors per item being .54 for syllables and .62 for trigrams. Further, very large interindividual differences in learning rate were observed. The recognition–learning model of Kintsch (1966) was used to describe the sequences of correct and incorrect recognitions. Because of the lack of homogeneity, the quantitative fit of the model was not very good. The predicted distributions of the total number of errors per item did not differ significantly (1%) from the observed distributions by the Kolmogorov-Smirnov one-sample test for each S. However, the predicted trial of the last error was too high for three of the Ss. Qualitatively the data were in agreement with the model, in that the prediction of a stationary error proportion on trials before the last error was confirmed. When all sequences before the last error were halved, the overall proportion of errors was found to be .39 in each half. Thus the basic notion of the model, that response probability suddenly increases to 1 from some constant initial level, is supported by the data.

The Ss apparently used confidence judgments consistently, since the proportion of correct responses is highest for the rating category 5 (very sure) and then decreases monotonically to category 1 (pure guess). Let an item which is presented with the correct response alternative be designated as an S-R item and an incorrect pairing an S-R′ item. An indication that Ss can discriminate very well between correct items and distractors on trials before the last error is obtained by computing the proportion of S-R items in each of the ten rating categories, from Yes-5 to No-5. The resulting values decrease in almost linear fashion from .76 to .22, showing an appreciable amount of discrimination on trials before the last error (Fig. 1). Note also that the posterior probabilities for the yes and no responses are almost the mirror image of each other. This is a peculiarity of the present experimental situation, since S was not only learning to say yes to S-R-type items but also to say no to S-R′-type items.

The confidence judgments were also used to construct MOCs. The technique of constructing MOCs from confidence ratings

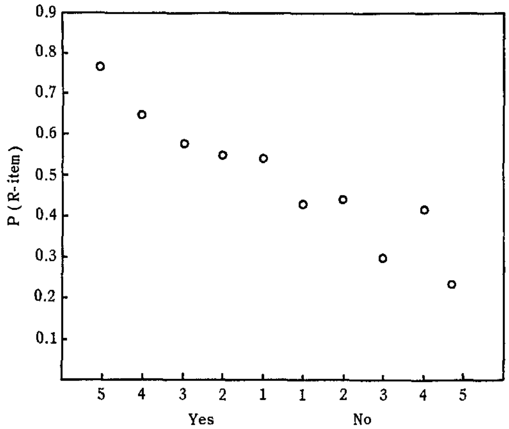

FIG. 1. Proportion of S-R-type items in each response class.

need not be described here in detail. Briefly, it consists of computing the proportion of S-R items which have been assigned to each of the ten response classes (from Yes-5 to No-5). When these proportions are cumulated ten different hit rates, R_i, are obtained, ranging from an extremely strict criterion (Yes-5) to the criterion actually employed by S (Yes-1), and to the laxest criterion (No-5). Similar computations are performed on the S-R′ items, resulting in the cumulated proportions R_i', which correspond to the false-alarm rates of the detection experiment. The MOC consists, then, of a plot of R_i vs. R_i'. Figure 2 shows the MOCs for all Ss for three different conditions. The curve in the upper left corner (open circles) is based upon all data from the first test trial. The data of primary interest have to do with S's performance on trials when, according to the Markov model, items have not yet been learned. These are the trials before the last error. Therefore MOCs were constructed from only those responses on the first test trial which were followed by an error on some later test trial. The full circles in Fig. 2 present the results of this computation. Two conclusions are evident from an inspection of Fig. 2: discrimination for items with a later error is clearly above chance, but it is substantially less than when all first-trial scores are considered. Finally, the triangles in Fig. 2 present the results of an analysis of the changes in discriminability on trials before the last error, except those of the first test trial. Thus the two MOCs (triangles and circles) are experimentally independent. In general the curves overlap almost completely; a small difference was obtained in only one case, but this S (LVM) was the fastest learner, and hence the data are based upon the fewest scores.

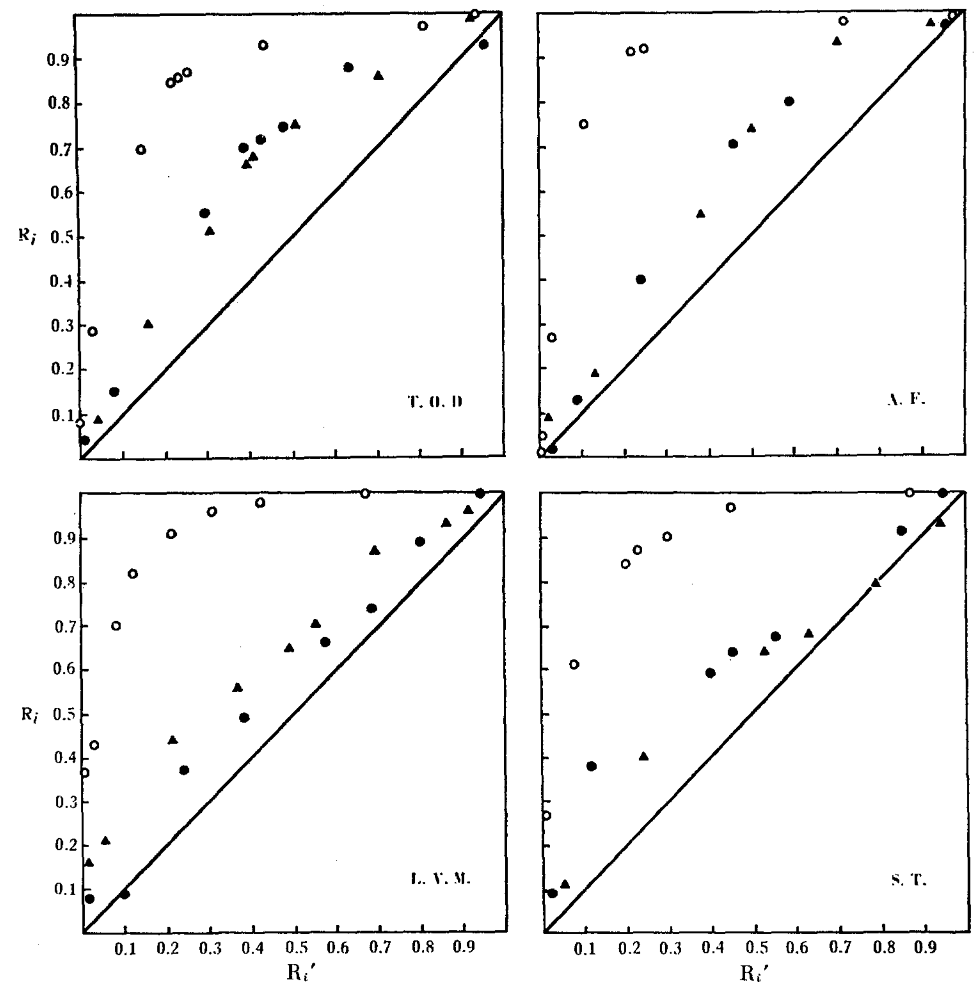

FIG. 2. MOC curves for all responses on the first trial (open circles), for first-trial responses followed by a later error (closed circles) and for responses after the first trial but before the last error (triangles).

We conclude from our data that discriminability did not change on trials before the last error. The data are, therefore, in agreement with the notion that the recognition-learning process involves a discontinuity.

While discriminability as measured by the MOC remains constant on trials before the last error, the confidence judgments themselves increase in a rather complex manner. All response protocols were rescored in eight different ways: assuming a very strict response criterion, i.e., calling a (correct) judgment of 5 "correct" and all other scores "error;" a less strict criterion by accepting as "correct" all correct 5's and 4's and letting all other responses be "errors;" and so on down to calling everything above incorrect-3 "correct," leaving only the incorrect responses with a confidence judgment of 4 or 5 as "errors." The proportion of "errors" in the first and second half of all trials before the last "error" was then computed for each set of transformed response protocols. Figure 3 shows the results averaged over all Ss, since the same conclusion holds for each one of them. A laxer criterion than correct-1 (the normal criterion) retains the stationarity on trials before the last error (none of the differences between halves were statistically significant). When the criterion is made stricter, performance on trials before the last "error" improves gradually: chi squares of 146.7, 99.6, 15.7, and 7.1 were obtained for the tests of stationarity when a criterion of 5, 4, 3, or 2 was used, respectively. Apparently confidence keeps increasing after learning has occurred. The implications of this finding are not completely clear. It might be regarded as evidence for a gradual over-learning effect: associations might be established in an all-or-none way, but once formed they are strengthened by further learning trials. A similar argument has been advanced by Rock (1957). The results presented in Fig. 3 are, however, open

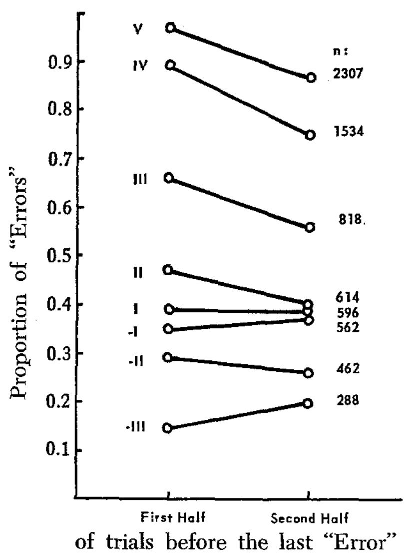

FIG. 3. Proportion of "errors" in the first and second half of trials before the last "error" for transformed scores. Roman numerals identify the transformations (see text). Arabic numerals refer to the number of cases in each half.

to several different interpretations, e.g., as indicating merely a general increase in confidence after learning, and no further use will be made of them.

Discussion

The two main results of this study are that: discriminability is greater than zero even before learning is completed, and discriminability does not change on trials before the last error. In order to be compatible with the present data a model of recognition learning must fulfill at least the following three conditions. First, it must describe the Markovian character of the response sequence: the probability of entering the learning state does not change over trials, and the recognition probability on presolution trials is constant. Second, it must provide some mechanism which can generate MOCs like the ones obtained here. Third, a model must be able to account for the relationship between the

MOC based upon all first-trial scores and the MOC based only on those scores which were later followed by an error. The present data do not provide enough information to specify all details of a satisfactory recognition learning model. However, some instructive suggestions as to the general outlines of such a model may be obtained.

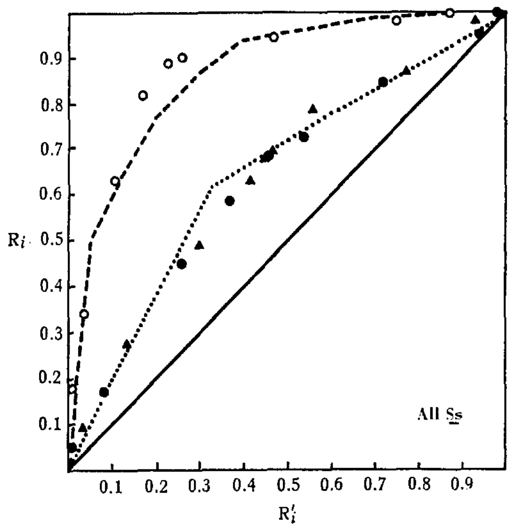

FIG. 4. Predicted and observed values for the average MOC curves; the conditions are the same as in Fig. 2.

It is possible to construct a model which is in at least qualitative agreement with the data by assuming that Ss are able to order items on a scale of oldness, or familiarity. The scale values of items seen for the first time are represented by a probability distribution $f_n(x)$; the scale value of repeated items have a probability distribution $f_o(x)$, the mean of which is higher than the mean scale value of the new items. Learning and forgetting thus consist in moving an item from one distribution to the other, in accord with the model of Norman and Wickelgren (1965). One must further assume that Ss use a second more stringent criterion to regulate the storage-process itself. Items with a lesser oldness value are subject to forgetting and, also, will be processed like new items on the next learning trial. Items with a greater scale value are permanently learned. These assumptions make the model equivalent to the recognition–learning model of Kintsch (1966), as far as the sequence of correct and incorrect responses during recognition learning is concerned. Figure 4 shows some predictions derived from this model. All parameters were estimated from the learning data only, without consideration of the confidence judgments. Because learning rates were not homogeneous in the present experiment, a serious quantitative test of the model can not be performed. Figure 4 is therefore presented merely as a demonstration that a theory is feasible which handles both the Markovian properties of the data and the MOCs.

REFERENCES

EGAN, J. P. Recognition memory and the operating characteristic. Technical Note AFCRC–TN–58–51, Indiana Univ., Hearing and Communication Laboratory, 1958.

KINTSCH, W. Recognition learning as a function of the length of the retention interval and changes in the retention interval. *J. math. Psychol.*, 1966, 3, 412–433.

MUDOCK, B. B. JR. Signal detection and short-term memory. *J. exp. Psychol.*, 1965, **70**, 443–447.

NORMAN, D. A., AND WICKELGREN, W. A. Short-term recognition memory for single digits and pairs of digits. *J. exp. Psychol.*, 1965, **70**, 470–489.

ROCK, I. The role of repetition in associative learning. *Amer. J. Psychol.*, 1957, **70**, 186–193.

(Received April 13, 1966)