Q. Jl exp. Psychol. (1969) **21**, 214–224

RECOGNITION VS. RECALL: STORAGE OR RETRIEVAL DIFFERENCES?

RICHARD D. FREUND, JOHN W. BRELSFORD, JR.† AND RICHARD C. ATKINSON

Stanford University, Stanford, California 94305

Differences between recognition and recall performance may be due to differences in storage processes, differences in retrieval processes, or some combination of both. An attempt was made to determine which process was critical by withholding information, at the time of study of a stimulus-response pair, about how that item was to be tested on its next presentation. It was found that differences between recognition and recall did not depend upon whether or not the subject knew, at time of study, the mode of test to be employed. These results were interpreted as support for the assertion that, in this particular task, differences in retrieval processes were sufficient to account for differences in recognition and recall. It was found that both the direction and magnitude of the recognition-recall difference depended upon the guessing correction employed.

Introduction

Differences in recognition and recall performance may be attributed to differences in storage processes, differences in retrieval processes, or some combination of both (Atkinson and Shiffrin, 1968). The present experiment is an attempt to determine whether recognition-recall differences may be observed in the absence of differential storage processes.

Consider the learning of a list of paired associates, where the set of response alternatives is well known to the subject. The study of a stimulus-response pair will be assumed to generate some input **I** to the memory system. Storage processes may be represented by an operator **S** applied to this input. The information stored in memory, denoted by **I′**, is therefore a function **S** of the input:

$$\mathbf{I}' = \mathbf{S}(\mathbf{I}).$$

At the time of test a retrieval operator **R** is applied to this stored information, which results in output **O**:

$$\mathbf{O} = \mathbf{R}(\mathbf{I}').$$

In short, the output is a function **R** of what is stored, which is in turn a function **S** of the input:

$$\mathbf{O} = \mathbf{R}(\mathbf{S}(\mathbf{I})).$$

† Present address: Yale University, New Haven, Conn., U.S.A.

This functional representation of memory permits a logical analysis of the various operations involved in recognition and recall. Typically the subject, throughout an experimental session, is aware of which mode of test is to be employed. Thus there is ample opportunity for utilization of different storage operations and different retrieval operations. If recognition and recall differ in storage processes, then what is crucial to the differential storage is knowledge of the mode of test to be employed. Without this knowledge there cannot be differential storage, and hence any performance differences may be attributed to retrieval processes alone.

In the present experiment, a list of paired associates was learned using an anticipation method. The procedure varied from the typical one, however, for any item could be tested either by recall or recognition on any trial, i.e. when an item was studied on trial n the learner did not know how it was to be tested on trial $n + 1$. Thus there was no opportunity for storage to take place in one way in anticipation of a recall test, and in another way in anticipation of a recognition test. Any observed differences between recognition and recall performance therefore would be attributable to different retrieval processes. Two control lists were also used in the experiment, one learned employing recall tests throughout, and the other, recognition tests. For these lists the subject knew at all times how an item was to be tested.

Method

A within-subjects design was employed, with each subject tested under all experimental conditions in each experimental session. The two independent variables were: (1) method of testing—recall or recognition; and (2) whether or not the subject knew, when studying a given stimulus-response pair, how that pair was to be tested.

Stimuli and responses

The stimuli were 432 three-letter nonsense syllables (Consonant-Vowel-Consonant pronounceable trigrams) of approximately 20 to 50 per cent association values. They were randomly selected from the Glaze list in Appendix A of Underwood and Schulz (1960). Eight different lists of 54 stimuli each were constructed. Each subject received a different list on each session. Throughout the experiment the responses were the digits 1 to 9.

Apparatus

Programming, generation of stimuli, and response recording were controlled by an on-line PDP-1 computer. Stimuli were electronically generated and displayed on the face of a cathode-ray tube (CRT). Responses were made on an electric typewriter located directly beneath the lower edge of the CRT. Each subject was seated in an individual sound-proofed experimental booth for each of his sessions.

Procedure

An anticipatory paired-associate learning procedure was used. The subjects were eleven Stanford students who received two dollars per session. Three to five practice sessions on other lists were completed for each subject to insure his understanding of the instructions. The experimental sessions were then initiated. No subject had fewer than five nor more than seven experimental sessions, and a total of 65-subject sessions were completed. For each subject-session one of the 54-item lists was selected at random. This list was subdivided into three 18-item lists. Responses were assigned randomly within each 18-item list, with the restriction that each response was assigned to exactly two stimuli.

The three 18-item lists can be distinguished by the type of retention test employed:

Recall (Re): One list of 18 items was always tested under recall conditions. On a test the stimulus would appear on the CRT, and the subject's task was to recall the response which was paired with the stimulus on prior study trials of the item.

Recognition (Ro): A second list of 18 items was always tested under recognition conditions. On a test the stimulus would appear on the CRT, along with the correct response and one incorrect (distractor) response. These response alternatives appeared to the right of the stimulus, one above the other, with the correct choice randomly located in the two positions. The subject's task was to choose the correct response for a given item from the two response alternatives displayed. The distractor was selected randomly on each trial for each recognition test.

Mixed: The third list of 18 items represented the crucial experimental manipulation. An item in this list could be tested either by recall or by recognition, the choice determined randomly at the time of test. Subjects were instructed that if the stimulus member appeared alone on the CRT they were to try to recall the correct response. If the stimulus member appeared with two response alternatives (displayed exactly as in the Ro condition) they were to try to recognize the correct response. Thus, a given item in this list could be tested on one trial for recall, but on the next trial could be tested either for recall or for recognition. This point was emphasized in the instructions. Those items from this "mixed" list which on any particular trial happened to be tested by recall shall be denoted as Recall Mixed (ReM) items and those tested by recognition as Recognition Mixed (RoM) items.

A "trial" consisted of one presentation of all 54 items. Within a trial the order of events was as follows: (1) A stimulus was selected from one of the three 18-item lists, e.g. from the Recall list, and was presented for test. The subject was given 3 sec. to respond; if he did not respond in this time period a message appeared on the CRT encouraging him to go faster. Next, the correct answer for that stimulus was displayed for 0·75 sec., followed by a 0·50 inter-item interval. (2) A stimulus from one of the remaining two lists, e.g. from the Mixed list, was selected and presented for test. Again the test period duration was 3 sec.,

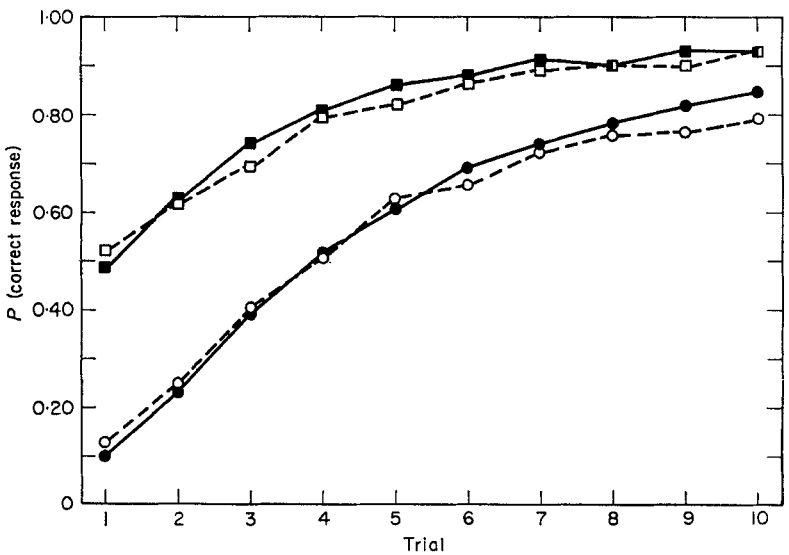

FIGURE 1. Mean proportion of correct responses as a function of trials for all conditions. —■—, Recognition; - - □ - -, recognition mixed; —●—, recall; - - ○ - -, recall mixed.

correct feedback was displayed for 0·75 sec., and inter-item interval was 0·50 sec. (3) A stimulus was selected from the remaining list (here, e.g. Recognition) for test and study just as for the two previous lists. Then this same order of selecting items would be repeated for three more items, and so on, until all 54 items had been tested. This represents trial 1. Each 18-item list was then instantly re-randomized and trial 2 began (without noticeable interruption). Ten such trials were completed in each session by each subject. Each day the subject received a new list of 54 stimulus-response pairs to learn. There are six permutations of the orders of testing of Recall, Recognition and Mixed; these permutations were assigned randomly among the 11 subjects (but remained fixed for each subject over all his sessions).

Crucial to the design is the fact that subjects knew what type of item was being presented at all times. To emphasize the distinction between the three types of items, they were displayed in a different third (top, centre or bottom) of the CRT. In addition, the appropriate label RECALL, RECOGNITION or MIXED appeared in capital letters to the far left of the stimulus (and remained on during the test period) for the three types of items.

The instructions explained the basic paired-associate learning task: learn the correct pairings between the CVC's and the digits. The three types of procedures were described in approximately the same fashion as in this report. Subjects were not told the number of items to be learned. They were instructed to make guesses (from number 1 to 9) if they could not recall the correct response. On recognition tests subjects were required to choose between the two response alternatives displayed on the CRT.

Results

The proportion of correct responses for each condition is plotted in Figure 1, averaged over all subjects and sessions. There are approximately 1400 observations at each point on the Recall (Re) and Recognition (Ro) curves, and 700 observations per point on the Recall Mixed (ReM) and Recognition Mixed (RoM) curves. It is clear that Recognition did not differ from Recognition Mixed, and that Recall did not differ from Recall Mixed. Analysis of variance was performed on the observed proportion of correct responses, summed over the ten trials

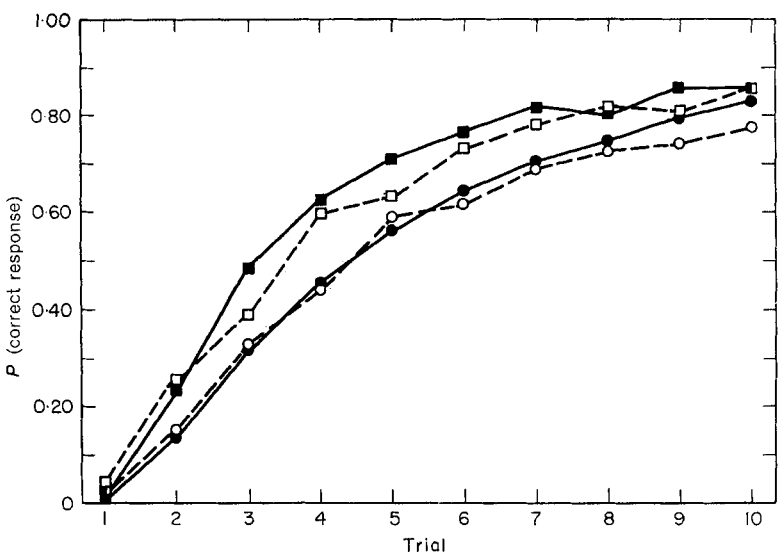

FIGURE 2. Mean proportion of correct responses as a function of trials, corrected for guessing For explanation of symbols, see legend to Fig. 1.

for each subject and condition, using the Newman–Keuls method (Winer, 1962, p. 114). Re vs. ReM and Ro vs. RoM did not approach significance. The four other pairwise comparisons (Re vs. Ro vs. RoM, Ro vs. ReM, ReM vs. RoM) were significant at the 0·01 level.

In order to compare recognition and recall performance some account must be taken of correct responses attributable to chance. A correction procedure frequently used (Hilgard, 1951, p. 556) involves the following transformation:

$$p' = p - \frac{1 - p}{N - 1},$$

where p denotes the observed proportion of correct responses in a given condition, p' the transformed proportion, and N the number of response alternatives (i.e. $N = 2$ for recognition tests, and $N = 9$ for recall tests). One interpretation of this transformation is that the observed proportion of correct responses is a weighted average of those items correctly retrieved from memory and those items correctly guessed.

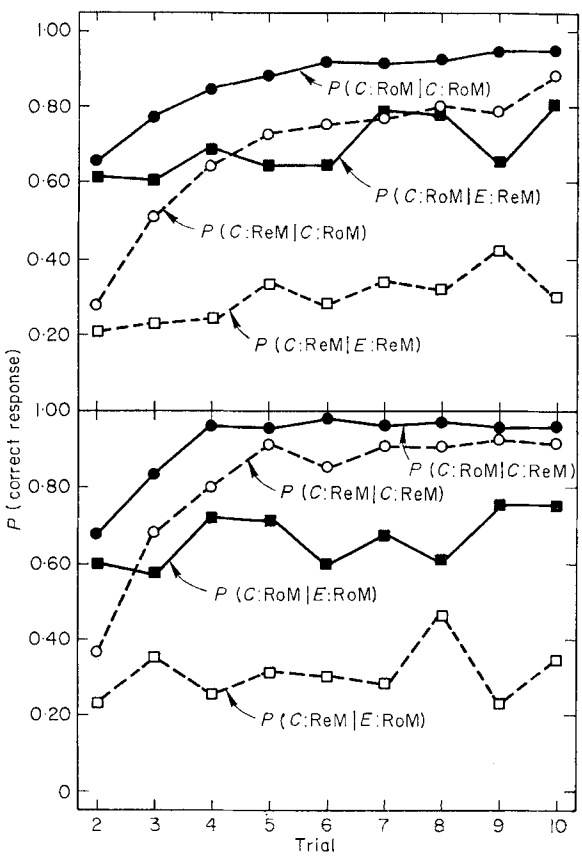

FIGURE 3. Mean proportion of correct responses as a function of trials for the Mixed Condition, conditionalized upon both the type of test and the response given on the preceding trial.

Application of this correction for guessing yielded the data plotted in Figure 2. Because of the correction for guessing all curves start at or near zero proportion correct. The learning of the recognition items appears faster than that of the recall items. The Newman–Keuls method of analysis of variance leads to rejection of the hypothesis that performance in the four conditions was equal ($F = 18{\cdot}48$, d.f. $= 3/30$, $P < 0{\cdot}01$). The six possible paired comparisons were then considered, and again it was found that Re vs. ReM and Ro vs. RoM did not approach significance. All other comparisons were highly significant ($P < 0{\cdot}01$). The reader should be alerted to the fact that these results are very much dependent upon the specific guessing correction employed; we shall return to this point later when we consider an alternative method for correcting for guessing.

An item in the mixed condition could be tested on any trial either by recall or by recognition. Figure 3 presents the proportion of correct responses on trial n for either type of test, conditionalized upon the mode of test on trial n-1 and also upon whether or not the response on trial n-1 was correct or an error. The legend code is $P(C{:}i \mid j{:}k)$, where $i =$ mode of test on trial n, $j =$ correct (C) or error (E) on trial n-1, and $k =$ mode of test on trial n-1. For example, in the upper panel of Figure 3 the uppermost curve, labelled $P(C{:}\text{RoM} \mid C{:}\text{RoM})$, represents the proportion of correct RoM responses on trial n, given a correct response on the previous RoM test trial for that item.

All curves in Figure 3 were tested for stationarity by Kendall's tau for multiple observations. The hypothesis of a constant (conditional) proportion correct over trials could be rejected for all but two of the curves at the 0·05 level of significance. The two stationary curves are both in the lower panel, namely, $P(C{:}\,\text{RoM} \mid E{:}\text{RoM})$ and $P(C{:}\text{ReM} \mid E{:}\text{RoM})$.

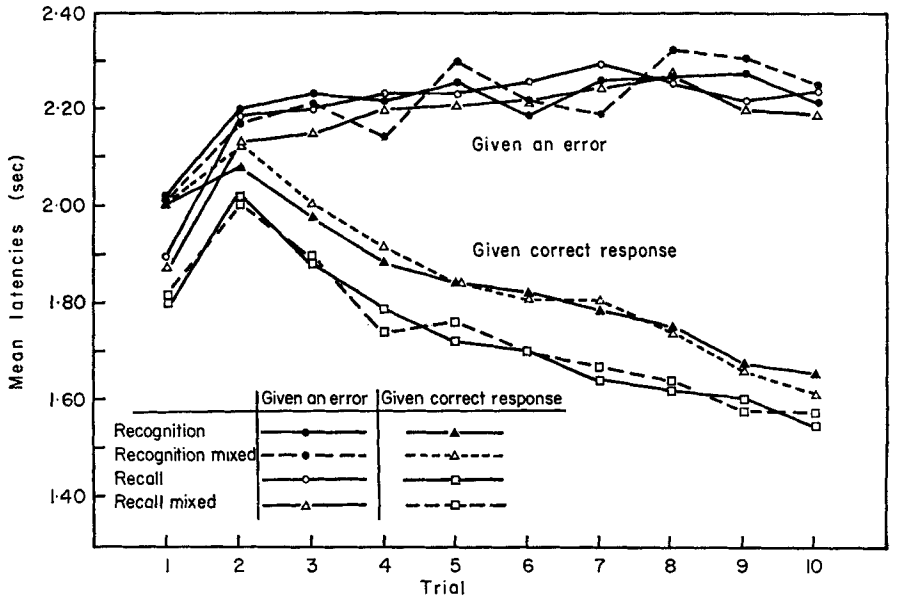

FIGURE 4. Mean latency of correct and incorrect responses as a function of trials for all conditions, averaged over subjects and sessions.

Latencies of responses conditionalized upon whether the response was correct or incorrect are presented in Figure 4. While latencies of incorrect responses remained fairly constant after the first trial (on which all responses were necessarily guesses), the latencies for correct responses showed a steady decrease over the course of the session. It appears that incorrect responses had equal latencies for all conditions, constant at approximately 2·2 sec. For correct responses, the two recall curves decreased together and the two recognition curves decreased together, but the latter were consistently about 0·1 sec. slower at each trial.

Discussion

The results indicate that the assumption of differential storage processes is not necessary to account for the differences between recognition and recall performance. The argument can best be presented by use of the simple equations introduced to relate input and output to the storage and retrieval operations. Recognition and recall tests may logically employ different storage and retrieval operations. Define $\mathbf{S}_{ro}$ and $\mathbf{R}_{ro}$ to be the storage and retrieval operators, respectively, for the recognition condition. Similarly, $\mathbf{S}_{re}$ and $\mathbf{R}_{re}$ are the storage and retrieval operators for recall conditions. To represent the storage processes in the mixed condition one more operator, $\mathbf{S}_{m}$, is required. Note that retrieval on the mixed tests utilizes either $\mathbf{R}_{ro}$ or $\mathbf{R}_{re}$. The output functions for the four conditions are therefore as follows:

$$\text{Ro:} \quad \mathbf{O} = \mathbf{R}_{ro}(\mathbf{S}_{ro}(\mathbf{I})) \qquad (1)$$

$$\text{RoM:} \quad \mathbf{O} = \mathbf{R}_{ro}(\mathbf{S}_{m}(\mathbf{I})) \qquad (2)$$

$$\text{ReM:} \quad \mathbf{O} = \mathbf{R}_{re}(\mathbf{S}_{m}(\mathbf{I})) \qquad (3)$$

$$\text{Re:} \quad \mathbf{O} = \mathbf{R}_{re}(\mathbf{S}_{re}(\mathbf{I})) \qquad (4)$$

Equations (1) and (2) show the relation between performance on the standard (i.e. where the test mode was known) and mixed recognition tests. Since only the storage operator is different, and since there were in fact no performance differences between these two conditions, we conclude that $\mathbf{S}_{ro} = \mathbf{S}_{m}$. The coincidence of the standard and mixed recall conditions leads to the conclusion, by inspection of equations (3) and (4), that $\mathbf{S}_{re} = \mathbf{S}_{m}$. Thus, all storage operators were identical.

The two mixed conditions, represented by equations (2) and (3), differ only in the retrieval operators applied to the stored information. It was between these conditions that the observed performance differences occurred. The conclusion is that differences between $\mathbf{R}_{ro}$ and $\mathbf{R}_{re}$ account for the performance differences obtained.

While the preceding analysis is logically consistent, it could be argued that neither differential storage nor differential retrieval took place in the present experiment. Rather, the superiority of recognition reflects only the greater "value" of identical information. For example, retrieval of the information, "The correct response is a high number," is quite valuable on a recognition test which presents 2 and 8 as response alternatives. Recall performance, given the same bit of partial information, would be inferior.

The existence of such partial information in the present experiment is suggested by the conditional analyses of Figure 3. From trial 4 onward it is clear that

correctly recalling a response implied virtually perfect recognition performance on the following trial. Conversely, failure to recognize an item led to near chance performance on the following trial for either test mode.

The above mentioned argument against either storage or retrieval differences clearly employs a more limited use of the term "retrieval" than does the present paper. Shiffrin and Atkinson (1969) have classified retrieval into three primary mechanisms: search, recovery, and response generation. Search is the process of locating the stored information in memory. Recovery is the process by which some or all of the information is made available for the generation of a response. This response generation is considered to comprise all aspects of translating the recovered information into the desired response. These aspects include the development of guessing strategies adopted within the context of particular test requirements. In other words, retrieval (as used here) includes the differential utilization of partial information, and therefore the differences obtained may appropriately be attributed to "retrieval" processes.

The finding that the conditional latencies for errors remain relatively constant while those for correct responses decrease over trials, is the typical result for paired-associate studies of this type (cf. Bower, 1967). The coincidence of the two recognition conditions, and that of the two recall conditions, lends further support to the hypothesis that identical processes took place regardless of whether or not subjects knew how they were to be tested.

More noteworthy is the fact that the correct recognition responses were consistently slower, from trial 2 onward, than those for correct recall responses. This may be explained in part by the following model, which assumes that latencies for responses retrieved correctly from memory have a fixed mean value L, whereas a retrieval failure and subsequent guess has a fixed mean value L'. Thus error responses always have a mean latency L'. Correct responses may occur as a result of a retrieval from memory or a correct guess, and consequently the latencies for correct responses represent a weighting of L and L'.

This model would be expected to have fair success for the following reason: a higher proportion of the correct responses in the recognition conditions were due to correct guessing. The relative weight of L' is therefore higher for recognition, and the latencies for corrects would be expected to be longer. An attempt to derive the quantitative predictions of the latency results, however, from this weighting function in conjunction with the observed probabilities of correct responses at each trial, revealed that this model is not adequate.

On correction of raw scores for chance successes. The success of the present experiment in the separation of storage and retrieval is dependent upon the existence of recognition-recall performance differences. The establishment of such differences raises the problem of correcting raw scores for guessing. The most common transformation is the one employed in this report. An alternative transformation of the raw scores is provided by the theory of signal detectability (TSD). Murdock (1966) has discussed the application of TSD to evaluation of recall and recognition performance. Table II in Elliott (1964, pp. 682–683) allows direct conversion of observed proportion of correct responses to d' as a function of the number of alternative responses.

The raw data of Figure 1 were transformed to d' scores and are replotted in Figure 5. The 9-choice values may be considered as the analogue of the present experiment's recall task if we ignore the fact of the physical absence of the 9 alternatives at time of test. The 9-choice values were obtained by plotting the theoretical 9-choice function on normal-normal co-ordinates, estimating the slope and y-intercept, and applying the linear approximation formula given by Elliott (1964, p. 680).

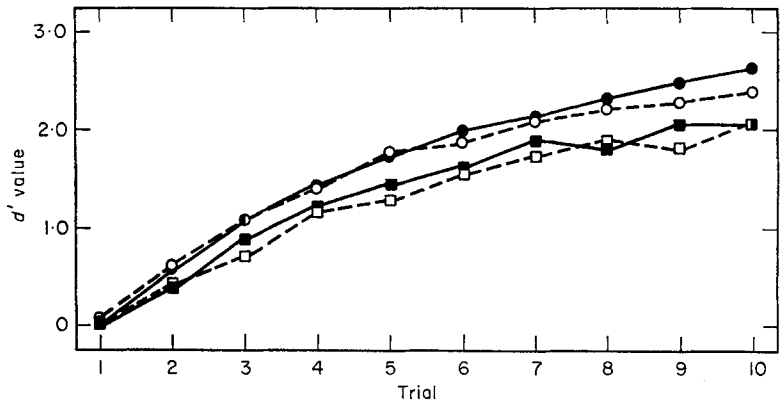

FIGURE 5. Mean values of d' as a function of trials for all conditions, averaged over subjects and sessions. For explanation of symbols, see legend to Figure 1.

While it was expected that the d' transformation might attenuate the superiority of recognition over recall, the completely surprising result obtained was a reversal of the direction of superiority: recall was now superior to recognition. To explain how this could come about, Figure 6 presents the theoretical results of

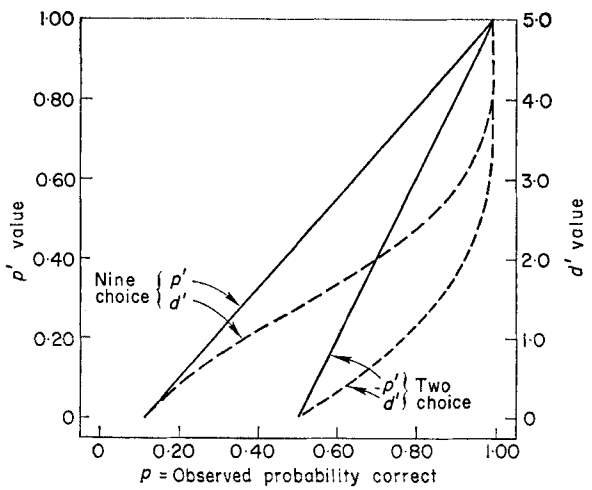

FIGURE 6. Values of p' and d' as a function of the observed probability of a correct response p, for the 2-choice and 9-choice tasks.

transforming 2 and 9 choice data by the standard correction for guessing (p') and TSD (d'). The former transformation yields the linear relation between the observed probability p and p'. The curved lines (to be read against the right-hand ordinate) show the d' value as a function of p.

Define a "reversal" as the event where one transformation of the data makes 2-choice superior to 9-choice and the other transformation yields the opposite ordering. An example of such a reversal occurs when the observed values are $p_2 = 0{\cdot}80$ and $p_9 = 0{\cdot}60$ (the subscripts indicate the number of alternatives). A p' transformation of these scores yields $p_2' = 0{\cdot}60$ which is greater than $p_9' = 0{\cdot}55$, whereas a d' transformation of the same scores yields a $d_2' = 1{\cdot}19$ which is less than $d_9' = 1{\cdot}70$. Hence the same observed scores yield opposite ordering; the p' transformation indicates that recognition is superior to recall, whereas d' suggests the opposite conclusion.

Figure 7 plots the entire space of all possible observations of p_2 and p_9. The two functions plotted may be called iso-probability curves: they are the locus of values

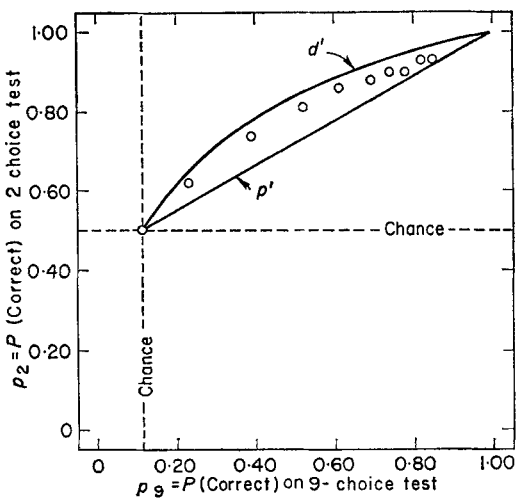

FIGURE 7. Relationship between p' and d'. (When the observed data point falls below the linear function, both the p' and d' corrections indicate that recall is better than recognition; for a data point above the bowed function both yield recognition as superior to recall. In the middle region, the p' and d' corrections yield opposite conclusions.) ○, Data observed on recall and recognition.

of p_2 and p_9 which must be observed to yield $p_2' = p_9'$ (for the linear function) and $d_2' = d_9'$ (for the bowed function). If both 2-choice and 9-choice performance are at chance (i.e. $p_2 = 0{\cdot}50$ and $p_9 = 0{\cdot}111$) then by either transformation the two performances will be judged equal and therefore the two curves intersect; the same is true for the case where $p_2 = p_9 = 1{\cdot}00$.

Consider now some other points in this space. If the observed data point (p_2, p_9) lies above the linear functions then using the p' transformation we conclude that recognition is better than recall (namely, that $p_2' > p_9'$). However, using the d' transformation, the observed data point (p_2, p_9) must fall above the bowed function for us to reach the same conclusion (namely, that $d_2' > d_9'$). Hence, any

data point bounded between the two functions will yield a reversal: the p' correction indicating that recognition is superior to recall, whereas the d' correction yields the opposite conclusion. The data points for the present experiment (the observed values of p_2 and p_9 on each trial for the Re and Ro conditions) are plotted in Figure 7, and as expected, lie in this middle region.

It is clear that the question of whether or not recognition is superior to recall remains meaningless until the correction for guessing is specified and defended. The defence of any such correction rests upon an assumption: namely, that the correction employed is reasonable with respect to the stated theory about the nature of the memory trace. It is maintained, on the basis of the preceding analysis, that special care must be given to the support of such assumptions before recognition and recall can be compared in any meaningful way.

Support for this research was provided by the National Aeronautics and Space Administration, Grant NGR-05-020-244.

References

ATKINSON, R. C. and SHIFFRIN, R. M. (1968). Human memory: A proposed system and its control processes. In SPENCE, K. W. and SPENCE, J. T. (Eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 2, pp. 89–195. New York: Academic Press.

BOWER, G. H. (1967). A descriptive theory of memory. In KIMBLE, D. P. (Ed.), *The Proceedings of the Second Conference on Learning, Remembering and Forgetting*, pp. 112–85. New York: New York Academy of Sciences.

ELLIOTT, P. B. (1964). Tables of d'. In SWET, J. A. (Ed.), *Signal Detection and Recognition by Human Observers*, pp. 651–84. New York: Wiley.

HILGARD, E. R. (1951). Methods and procedures in the study of learning. In STEVENS, S. S. (Ed.) *Handbook of Experimental Psychology*, pp. 517–67. New York: Wiley.

MURDOCK, B. B. (1966). The criterion problem in short-term memory. *J. exp. Psychol.* **72**, 317–24.

SHIFFRIN, R. M. and ATKINSON, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychol. Rev.* **76**, 179–193.

UNDERWOOD, B. J. and SCHULZ, R. W. (1960). *Meaningfulness and Verbal Learning*. New York: Lippincott.

WINER, B. J. (1962). *Statistical Principles in Experimental Design*. New York: McGraw-Hill.

Received 22 February 1969