Confidence ratings in continuous paired-associate learning*

HARLEY A. BERNBACH, Cornell University, Ithaca, N.Y. 14850
and
GORDON H. BOWER, Stanford University, Stanford, Calif. 94305

Confidence ratings were collected in a continuous paired-associate learning task in which items were presented three times each. Analysis of Type 2 operating characteristics showed no difference in the discriminability of correct responses from errors after one vs two reinforcements. Increasing confidence ratings across trials were attributable to a shift in criterion.

It is commonly known that confidence ratings in recall experiments increase under conditions in which the probability of a correct response increases. Thus one would expect an increase in the average response judgment across trials in a paired-associate learning (PAL) experiment. Evidence for this phenomenon (e.g., Suboski, Pappas, & Murray, 1966) has generally been taken as evidence against an all-or-none position regarding the nature of the memory trace.

Murdock (1966) constructed Type 2, or response-conditional, operating characteristics (OCs) from confidence ratings in a short-term paired-associate probe experiment. He found that the detectability index, d′, did not vary with serial position, despite a marked variation in the probability of a correct response. He showed, however, that a criterion shift accompanied each change in response probability and that the average confidence ratings reflected these criterion shifts. That is, Ss usually gave higher confidence ratings to responses recalled to items from later serial positions, although they were unable to discriminate correct from incorrect responses any better for these positions than for earlier ones.

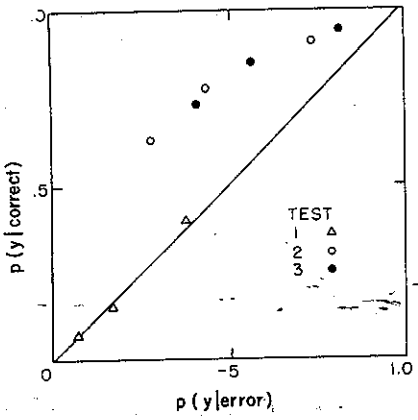

Fig. 1. Group ROC curves for three values of d′.

The independence of d′ from memory factors influencing response probability, such as serial position, was taken by Bernbach (1967) as evidence for the all-or-none nature of the memory trace. He argued that a theory relating confidence judgments and response probability through a common intervening variable, such as habit strength, would predict an increase in the Type 2 d′ measure with increases in recall probability. On the other hand, his finite-state decision theory, having one distribution of confidence ratings for remembered items and another for items not remembered, predicts no effect of memory factors on d′. Bernbach (1967) presented evidence from replications of the Murdock (1966) probe study and from standard PAL studies that supported this finite-state assumption.

The purpose of the present experiment was to investigate the relationship between confidence judgments and recall probability in the continuous PAL task. In addition, an attempt was made to investigate whether the meaningfulness (M) of the S-R pair affected this relationship.

METHOD

Ss were 18 university students who were paid for their participation. Each S learned a continuous list of 380 consonant bigram-consonant pairs by the anticipation method. As each consonant bigram was presented visually, S anticipated the correct consonant response (from the set B, C, F, H, L, P, S, and T) by pressing a button labeled with that response. He then indicated his confidence that his response was correct by pressing one of four buttons marked from least to highest confidence. The confidence response was followed by display of the entire stimulus-response pair. The S-R pair when shown together may be considered as a CCC trigram, and a range of *meaningfulness* values (Witmer, 1935) was used over the different trigrams S learned.

The list of 380 items contained 117 critical items that appeared three times each, with six other items intervening between each presentation. Warm-up effects were minimized by presenting 11 filler items before the first presentation of a critical item.

RESULTS AND DISCUSSION

As expected, both response probability and the mean confidence ratings increased with presentations, as shown in Table 1. The mean confidence ratings were determined by assigning the numbers 1 to 4 to the confidence responses, ranging from least to most confident.

The ability of the Ss to discriminate their correct responses from their errors is shown graphically by the OCs in Fig. 1. On the first test, before the stimulus-response pair had been presented for study, no discrimination was possible and the OC falls on the chance line. On Tests 2 and 3, however, the OCs indicate that Ss could discriminate correct from incorrect responses, but could do so no better on Test 3 than on Test 2. Like the data reviewed by Bernbach (1967), these results are consistent with a two-state theory of confidence judgments and recall.

The data in Fig. 1 also show the criterion shift noted by Murdock (1966). That is, the points for Test 3 are moved up and to the right from those for Test 2, though they lie on the same OC. Ss were apparently able to extract some information about the presentation number from the stimulus and use this information to lower their criterion for later presentations. Therefore, the relation between average confidence and presentation number that is shown in Table 1 apparently results from a bias to give higher ratings to later presentations of items, independently of whether or not the item is correctly recalled. In other words, it does not indicate a direct relation between recall probability and confidence.

Table 1
Response Probability and Mean Confidence Ratings

Presentations	Proportion Correct Responses	Mean Confidence Rating
1	.175	1.90
2	.385	2.78
3	.457	3.13

*This research was supported by Grant MH 13950 to the second author from the National Institute of Mental Health. The work was performed while the first author was a National Science Foundation postdoctoral fellow at Stanford University.

Unfortunately, our attempt to investigate the effect of M on these relationships was not successful, as we found no effect of M of the S-R pair on recall probability in this experiment. Treating as M the association value (AV) of the stimulus-response pair as a trigram (Witmer, 1935), the correlation of proportion correct and AV was only .02. We attempted further to correlate proportion correct on an item with the normative frequency with which the third letter or S-term (Underwood & Schulz, 1960, p. 374). Despite wide variation of this measure in our sample, it accounted for practically none of the variation across items in the proportion correct ($r = .11$, $p > .10$).

REFERENCES

BERNBACH, H. A. Decision processes in memory. Psychological Review, 1967, 74, 462-480.

MURDOCK, B. B., JR. The criterion problem in short-term memory. Journal of Experimental Psychology, 1966, 72, 317-324.

SUBOSKI, M. D., PAPPAS, B. B., & MURRAY, D. J. Confidence ratings in recall paired-associates learning. Psychonomic Science, 1966, 5, 147-148.

UNDERWOOD, B. J., & SCHULZ, R. W. *Meaningfulness and verbal learning.* New York: Lippincott, 1960.

WITMER, L. R. The association value of three-place consonant syllables. Journal of Genetic Psychology, 1935, 47, 337-360.

Recency/primacy ratio: A short test of task orientation

PETER HAMILTON
Medical Research Council Applied Psychology Unit, Cambridge, England
and
G. R. J. HOCKEY
University of Durham, Durham, England

Thirteen Ss recalled 60 nine-digit lists presented consecutively for 15 min at a rate of two digits/second. An analysis of errors by serial position revealed that (1) primacy errors in the first two list positions increased over time, and (2) recency errors in the last two list positions decreased over time. The ratio recency errors/primacy errors showed a consistent downward trend over the testing period. It is suggested that this ratio is an index of the degree of active information processing and may be a useful reflection of the vigilance type of decrement, which can be derived from the use of a short simple test.

The measurement of performance of the individual under conditions of stress, prolonged work, and diverse affective states is often frustrated by the lack of sensitive tests. A major criticism seems to be task duration, only relatively long tasks being successful in reflecting such changes in performance as occur under these conditions. Broadbent (1958) and Wilkinson (1968), for example, list the desirable characteristics of tests for measuring effects of noise and sleep loss, respectively; task duration is regarded as of primary importance in both cases. Short tests are generally not sensitive to such changes in the state of the individual as can readily be achieved in the laboratory, though, in real-life situations (such as clinical diagnoses and industrial measurement), the considerable time pressures involved make the development of short sensitive tests highly desirable.

Two broad categories of "traditional" tests of such performance changes can be distinguished. (1) Perceptual-motor, "information-throughput" tasks, such as tracking, and the five-choice serial reaction task (Broadbent, 1963); performance decrement in these tests appears as an increase in either the variance or the length of response times over a ½ h or so, or as an increase in tracking error, such as time-off-target. (2) Vigilance tests (Davies & Tune, 1970); here, degradation takes the form of an accelerated decrement in detection probability over time or of a general lowering of detection probability. Because of the wide range of individual variation on both kinds of test, a repeated-testing design is usually necessary to establish reliability.

The present experiment is motivated by the need for simple "one-shot" tests of performance, applicable to group testing situations. The minimum criteria for such a test are that it is sensitive to the typical decrement with time on task, in a short time period, and in a single testing session. An additional desirable aspect of such a test is that it should be portable and simple to administer (the kinds of tasks described above often require considerable practice because of the unusual nature of the task demands).

Auditory short-term memory, in which 8-10 digits or letters are presented sequentially, possesses the basic qualities of such a test, though, as yet, it has not been used to study performance change over a period of repeated presentations. Performance on such tests is characterized by a pronounced serial-position curve of error. Various theories (Broadbent, 1958; Waugh & Norman, 1965; Glanzer & Cunitz, 1966) have attributed the shape of the curve to two components: one, a very short-term "echo-box" store with a time constant of a few seconds, which ensures efficient recall of the last few items (recency); the other, a longer-term store which is assumed to be the source of high recall of the first few items input (primacy). It is likely that the extent of the primacy effect is dependent on the rate of information processing during the input of the list. Crowder (1969) found that the primacy effect disappeared on the standard nine-digit list he was studying when he presented Ss with a series of lists of different length. He attributes this effect to lack of active rehearsal of the first few items—the S anticipating a much longer list.

If the reduction in active information processing seen in the Crowder paradigm is comparable with that occurring when decrement is observed in perceptual-motor tasks or vigilance situations, we should anticipate a loss of primacy in a condition that requires continuous memorizing and recall of short lists. Moreover, the amount of data to be obtained in a short period with such a test should ensure a high degree of reliability within an economical experimental paradigm.

SUBJECTS

Thirteen adult males and females were paid to attend the session.

DESIGN AND PROCEDURE

Ss were presented with 60 nine-digit lists and asked to recall each immediately after it occurred. Within each list, digits were presented at two/second, synchronized with metronome beats. An interval of

Table 1
Mean Percent Error at Each List Position

Position	1	2	3	4	5	6	7	8	9
Percent Error	5.0	22.4	38.7	46.0	58.5	66.7	65.9	50.3	13.5

Table 2
Mean Percent Correct Recall Over Six Blocks of 10 Lists Each

Block	1	2	3	4	5	6
Percent Correct	58.0	60.0	56.7	57.5	62.9	60.7