Journal of Experimental Psychology
1969, Vol. 81, No. 1, 126–131

DISCRIMINATION OF ITEM STRENGTH AT TIME OF PRESENTATION[1]

TANNIS Y. ARBUCKLE[2] AND LOLA L. CUDDY

Sir George Williams University *Queen's University*

Two experiments examined *S*s' ability to discriminate differences in associative strength of paired-associate items. In both experiments, lists of S–R pairs were presented for single study-test trials. As each S–R pair was shown, *S* predicted whether he would recall R correctly. In Exp. I, accuracy of "yes"–"no" predictions of practiced *S*s was significantly greater than chance expectations, while in Exp. II, a reliable decrease was found in probability of correct recall as predictions of naive *S*s varied on a scale from "very likely" to "very unlikely" (to recall). Accuracy of *S*s' predictions appeared related to an ability to assess item difficulty rather than to use of special learning strategies. It was concluded that items do differ in associative strength immediately following presentation and that *S*s can discriminate these differences.

When recall of any paired-associate (PA) item is tested immediately after presentation, probability of correct recall is near 1.0 (e.g., Tulving & Arbuckle, 1963). As Underwood (1964) pointed out, the fact that all items can be recalled means that some measure other than recall probability is needed to assess the possibility that items differ in associative strength at time of presentation.

In the present experiments it was argued that if PA items differed in associative strengths immediately following presentation, *S*s should be able to detect these differences just as they can detect differences in strength of any other form of input signal. Specifically, *S*s were asked, as each pair was presented for study, to predict whether or not they would recall it, the assumption being that any tendency for the accuracy of their predictions to exceed chance expectations would indicate both that differences in associative strength of PA items were present immediately following presentation and that *S*s could respond to such differences.

[1] Part of this paper was presented at the meeting of the Canadian Psychological Association, Calgary, 1968. The research was supported by Grant APA 209 to the first author from the National Research Council of Canada. The authors thank David Baxter, Tonnar Brace, and Margaret Zysman for assistance with data collection in Exp. I and Frances Battista for assistance with data collection and analysis in Exp. II.

[2] Requests for reprints should be sent to Tannis Y. Arbuckle, Department of Psychology, Sir George Williams University, Montreal 25, Canada.

Experiment I

Method

Subjects.—Two male and two female Sir George Williams University students participated as *S*s, while an additional 150 students served as judges of item difficulty. All *S*s and judges were paid and all had prior experience in PA learning.

Materials.—An item pool containing all possible pairings of all letters of the alphabet except I and O as S members with numbers from 10 to 99 inclusive as R members provided the PA items. Two sets of lists were prepared: Set A consisting of 60 six-item and 100 five-item lists, with S members from A–M and A–K, respectively; and Set B consisting of 60 six-item and 105 seven-item lists, with S members from N–Z and L–Z, respectively. Lists were constructed by randomly drawing items without replacement from appropriate sections of the pool, subject to the restriction that no S or R member was repeated within any list. The randomly drawn order, called Order 1, determined the presentation sequence of lists and of pairs within lists. A second order, Order 2, was constructed with the same list sequence and list membership as Order 1, but with different pair sequences within lists.

Additional 18 six-item lists of high-frequency word–word pairs (Thorndike-Lorge, 1944, count $\geqq 30$ per million) were constructed for use in preliminary training.

All S–R and S items were printed on 3×5 file cards.

Procedure.—On commencing the experiment, *S* was told that his main task was to try to learn all the PA items, but that in addition, as each pair was presented, he was to predict whether or not he would recall it. He was warned not to adopt any strategies which might artificially increase his accuracy of prediction.

Learning of a list consisted of a single study–test trial. As each pair was presented for study, *S* read

TABLE 1

Observed (O) and Expected (E) Frequencies for Four Combinations of Predicted Recall and Actual Response for Three List Lengths

L	S	Correct predictions				Incorrect predictions				χ^2
		P_cR_c		P_xR_x		P_cR_x		P_xR_c		
		O	E	O	E	O	E	O	E	
6	PA	200	186.9	38	24.9	90	103.1	32	45.1	12.3**
	DB	170	155.7	55	40.7	86	100.3	49	63.3	9.2*
	NS	107	69.3	151	113.3	48	85.7	54	91.7	63.4**
	MZ	84	50.6	173	139.6	63	96.4	40	73.4	55.0**
5	NS	162	114.7	183	135.7	72	119.3	83	130.3	70.5**
	MZ	92	58.7	245	211.7	54	87.3	109	142.3	43.3**
7	PA	444	421.8	58	35.8	181	203.2	52	74.2	23.0**
	DB	307	238.5	204	135.5	94	162.5	130	198.5	105.5**

* $p < .01$.
** $p < .001$.

it aloud and then predicted "yes," meaning he would be able to recall it correctly (P_c), or "no," meaning he would not be able to do so (P_x). In the test phase, S read aloud the S member, responded with either the correct two-digit number (R_c) or an incorrect two-digit number (R_x), and finally evaluated his answer as "right" (E_c) or "wrong" (E_x).[3] No feedback was given by E regarding the correctness of predictions, responses, or evaluations.

The sequence of presentation of pairs in the input phase was always different from the sequence of testing in the output phase. For each subset of L lists (where L = list length), there were L different output sequences, constructed using the method of Tulving and Arbuckle (1963) as modified by Arbuckle (1967).

Each S was tested individually in nine sessions spaced over 11 days. In Session 1, Ss learned the 18 practice lists; in each of Sessions 2–4, they learned 20 six-item lists; and in each of Sessions 5–9, the two Ss assigned to Set A learned 20 five-item lists, while the two Ss assigned to Set B learned 21 seven-item lists. The two Ss assigned to the Set A lists were the two who made the fewest correct responses in Session 1. This mode of assignment kept actual probability of correct recall across list lengths and Ss as near to .5 as possible, thereby maximizing for all Ss the difficulty of predicting recall.

The S–R and S cards of each list, plus two blank cards inserted at the beginning of the input and output sequences, respectively, were stacked in the appropriate order in a manually operated card holder attached to a plywood screen that separated E from S. All cards were successively removed at a constant 3-sec. rate paced by the beat of a timer, with removal of the first blank signaling the beginning of the trial and presentation of the second blank providing a 3-sec. interval between input and output phases.

On completion of the learning portion of the experiment, two new series of pairs were constructed. Series A, derived from Set A, contained the first 105 pairs from Sessions 5 and 6 for which the predictions were in agreement (i.e., the two Ss both predicted correct recall or both incorrect recall). Series B was constructed in identical fashion using 105 Set B pairs. The 150 judges were asked individually to rate these pairs on a 5-point scale going from 1—"easy to learn"—to 5—"hard to learn," 75 judges being assigned to rate each series of pairs.

Results and Discussion

The data of primary interest were the frequencies with which Ss correctly predicted correct and incorrect responses. Table 1 shows the observed frequencies of the two kinds of correctly predicted responses (P_cR_c and P_xR_x) and of the two kinds of incorrectly predicted responses (P_cR_x and P_xR_c), together with the expected frequencies of each under the hypothesis of no relation between prediction and recall. For all eight $S \times$ List Length conditions, predictions were significantly more accurate than expected ($p < .01$ in all cases). The actual percentages of accurately predicted responses, which ranged from 62% to 72% across the eight $S \times$ List Length conditions, were consistently lower than the percentages of accurate postresponse evaluations, which ranged from

[3] Postresponse evaluations were not made under the first two $S \times$ List Length conditions tested (DB–6; MZ–6).

84% to 92% across the six $S \times$ List Length conditions for which they were made. Nonetheless, there was clearly some relation between *S*s' predictions and their recall, and hence the data were analyzed further to try to determine both how *S*s arrived at their predictions and how the accuracy of predictions was affected by events occurring between prediction and attempted recall.

Bases of Ss' predictions.—One factor that seemed likely to influence *S*s' predictions was the apparent difficulty of the pair. Perceived item difficulty, as indicated by judges' ratings, seemed primarily determined by the association value of R members, with pairs containing numbers with high association values (Battig & Spera, 1962) being judged easier. When predictions of correct recall were looked at as a function of judged difficulty (see Fig. 1), it was found that relative frequency of P_c generally decreased as rated difficulty increased. (The number beside each data point is the number of items assigned that modal difficulty rating.) The two curves of Fig. 1 present two independent analyses of the relation between item difficulty and *S*s' predictions since items, judges, and *S*s were all different.

The difference in intercept between the two curves of Fig. 1 reflects marked individual differences among *S*s in frequency of prediction of correct recall. For each *S*, under each condition of list length, frequency of P_c was remarkably similar to frequency of R_c and, if made, of E_c, suggesting that possibly *S*s were using internal feedback regarding correctness of previous responses to determine how many items to assign to the P_c category.

It is always possible with the present design that *S*s' predictions are affected by use of rehearsal strategies, although the four *S*s here stated on postexperimental inquiry that they had neither stressed nor ignored any item. One rehearsal strategy that could have greatly biased the results, that of ignoring P_x items, was examined statistically. Specifically, if *S*s were ignoring P_x items, the proportion of P_xR_x responses that were half right (e.g., A–73 recalled as A–23 or

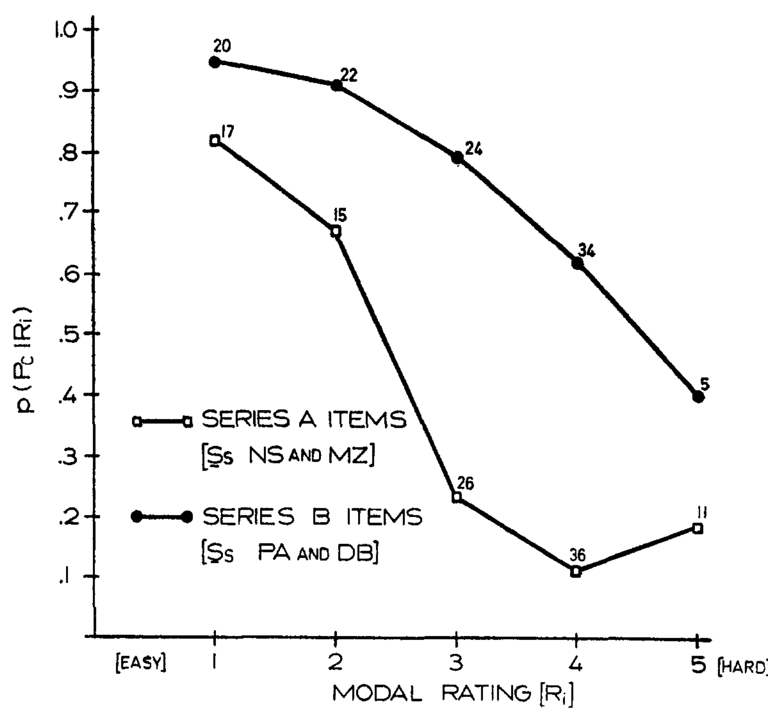

FIG. 1. Probability of predictions of correct recall as a function of rated item difficulty for two sets of 105 PA items.

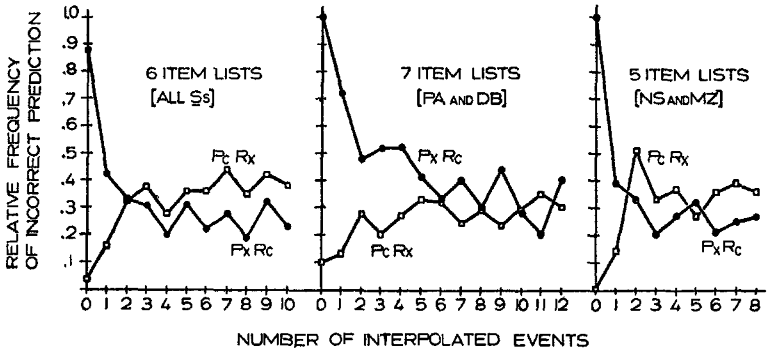

FIG. 2. Relative frequency of two types of incorrect prediction as a function of number of events interpolated between input and output of each item for three list lengths.

A–76) should approximate the chance expectation of .19. For the eight $S \times$ List Length conditions, the observed proportions ranged from .28 to .43 and were all significantly greater than chance ($z > 2.33$, $p < .01$, in all eight cases). Thus, it appeared that this particular rehearsal strategy was not being used.

Effects of intratrial events.—As Fig. 2 shows, with short retention intervals, P_c predictions were almost certain to be right and P_x ones to be wrong, but that after three or more interpolated events, both kinds of prediction were wrong about 30% of the time. Thus, while intratrial events did affect accuracy of prediction, the stable asymptotes of the curves in Fig. 2 ruled out any explanation of the bulk of the inaccurate predictions in terms of intratrial events.

Conclusion.—Experiment I indicated that practiced *S*s could predict their recall with above-chance accuracy. The question of whether the same results would be obtained with naive *S*s was examined in Exp. II.

EXPERIMENT II

Method

Subjects.—Forty Queen's University students, naive to PA learning, served as *S*s to complete a course requirement.

Materials.—Two sets, A and B, of 15 five-pair lists were constructed by randomly selecting, pairing, and assigning to lists 300 A-frequency nouns from the Thorndike-Lorge word count. All S–R and S items were handprinted on Graflex slides for presentation by a Carousel projector.

Design.—All *S*s were tested under two conditions using the recall-probe technique. Under Cond. PE, *S*s rated at time of presentation the probability that they could correctly recall each item, and at time of recall, the probability that they had correctly recalled the R member of the probe item. Under Cond. E, *S*s rated only at recall. The *S*s were tested in groups of 10, each group receiving the 15 lists of one set under one condition first, followed by the 15 lists of the alternative set under the other condition. Order of conditions and list sets was completely counterbalanced across groups of *S*s, and a different random sequence of pairs within lists was employed for each group. The serial position (SP) of the pair to be probed for recall was randomly determined with the restriction that all five SPs were probed three times over the 15 lists.

Procedure.—Under both conditions, *S*s were told that pairs would be presented at a 3-sec. rate, that they were to try to learn all pairs, and that they would be tested for recall of one pair.

In Cond. PE, *S*s were given a 15-page prediction booklet, the cover of which featured a PA item plus the question, "How sure are you that you will remember the pair?" Each page contained five 60-mm. horizontal lines, each with a 20-mm. bar at each end. The phrase "very likely" was printed to the left of each line and the phrase "very unlikely" to the right. The lines, sequentially from the top of the page, represented the five successive pairs of a list. The *S*s were told that as each pair was presented, they were to mark the corresponding line so that the location of the mark on the line indicated their prediction regarding the likelihood of their recalling the R member.

In both conditions, *S*s were given an answer sheet which provided for both the recording of responses and the evaluation of their correctness. At the top was printed "How sure are you that you are right?" Down the left side were 15 response blanks for *S* to enter the response to each of the 15 probes, and beside each blank was an evaluation line identical in dimensions and labels to those in the prediction booklet. The *S*s were instructed to use these lines to evaluate their confidence in the correctness of each response.

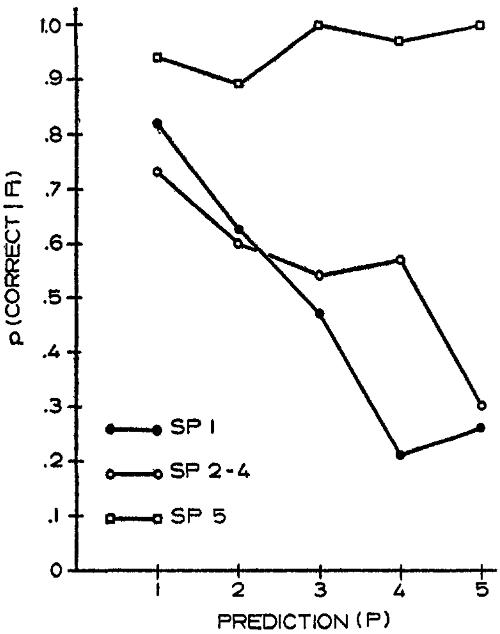

FIG. 3. Probability of correct response for each prediction for SP 1, SP 2–4, and SP 5.

Results

The correct recall data were analyzed for the effects of stage of practice (first or second list learned), list set (A or B), and condition (PE or E). The only significant effect was between conditions, with *S*s recalling correctly a mean of 9.40 items under Cond. PE and a mean of 8.26 items under Cond. E, $F\ (1, 36) = 6.22$, $p < .05$.

Predictions and evaluations were scored by dividing the rating line into 5 equal intervals, numbering these intervals from 1 to 5 from the "very likely" end of the line, and scoring each rating from 1 to 5 according to the interval marked. When predictions and evaluations were tabulated, predictions were found to be distributed fairly evenly over all ratings, while evaluations were restricted almost entirely to 1 and 5.

The data of primary interest, the conditional probabilities of correct response given the *i*th prediction, where $i = 1$ to 5, are shown in Fig. 3. For SP 1, and for SPs 2–4 averaged together, probability correct was highest for Prediction 1 and decreased in an orderly fashion across predictions. For SP 5, *S*s distributed their predictions among the 5 ratings, but were nearly always correct at recall; hence the function relating probability correct to prediction category was flat.

The question of whether the observed decrease in probability correct as predicted likelihood of recall decreased held true for individual *S*s was assessed by calculating, for each successive pair of predictions, the proportion of *S*s who correctly recalled more of the items given the *i*th prediction than of those given the $i + 1$th prediction. These proportions were .75, .70, .67, and .63 for the four comparisons where $i = 1$, 2, 3, and 4, respectively. A test of the probability of combined results (Maxwell, 1961), applied to these proportions, yielded a $\chi^2(8) = 24.34$, $p < .01$. This finding indicated that the observed relation between prediction and recall was reliable for individual *S*s and not merely a result of averaging.

The relation between recall and evaluation was examined for Evaluations 1 and 5 only since the other ratings were so little used. For Evaluation 1, the proportion of items correctly recalled was .94, and for Evaluation 5, it was .16.

General Discussion

The main finding was that both naive and practiced *S*s could reliably predict at time of presentation of an item whether they would recall it. One might argue that this finding resulted from rehearsal strategies, even though their use was controlled through instructions and variation of test orders in Exp. I and through use of the probe technique and of a rating continuum (as opposed to discrete categories) in Exp. II. While these controls seemed successful in that they prevented such obvious strategies as ignoring P_x items, it is not absolutely certain that they were completely successful. What is clear is that *S*s did not respond randomly to items. That item difficulty can be reliably discriminated and that it is predictive of learning was demonstrated for independent groups of *S*s by Underwood, Rehula, and Keppel (1962). The present experiments indicated that *S*s can use information contained in the item to predict their own learning. Since making predictions did not interfere with learning (as witness the superior recall under Cond. PE in Exp. II), one might speculate that the kinds of assessment of input information required overtly by the prediction task are in fact made covertly in standard PA learning situations.

The present experiments have essentially shown a correlation among associative strength of items, *S*s' predictions, and *S*s' recall. The obvious next step would be to seek out possible causal relations by manipulating discriminability of items along a dimension of associative strength and observing the effect on accuracy of *S*s' predictions under conditions where recall probability is held relatively constant.

REFERENCES

Arbuckle, T. Y. Differential retention of individual paired associates within an RTT "learning" trial. *Journal of Experimental Psychology*, 1967, **74**, 443–451.

Battig, W. F., & Spera, A. J. Rated association values of numbers from 0–100. *Journal of Verbal Learning and Verbal Behavior*, 1962, **1**, 200–202.

Maxwell, A. E. *Analysing qualitative data.* London: Methuen, 1961.

Thorndike, E. L., & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College, Columbia University, Bureau of Publications, 1944.

Tulving, E., & Arbuckle, T. Y. Sources of intratrial interference in immediate recall of paired associates. *Journal of Verbal Learning and Verbal Behavior*, 1963, **1**, 321–334.

Underwood, B. J. Degree of learning and the measurement of forgetting. *Journal of Verbal Learning and Verbal Behavior*, 1964, **2**, 112–129.

Underwood, B. J., Rehula, R., & Keppel, G. Item selection and paired-associate learning. *American Journal of Psychology*, 1962, **75**, 353–371.

(Received September 3, 1968)