Acta Psychologica 57 (1984) 215–235
North-Holland

RESPONSE LATENCY AND RESPONSE ACCURACY AS MEASURES OF MEMORY *

Colin M. MacLEOD
University of Toronto, Canada

Thomas O. NELSON
University of Washington, USA

Accepted January 1984

The relationship is examined between response accuracy and response latency as measures of memory, and questions are raised concerning the value of the unidimensionality assumption often invoked in theories of memory. Three paired-associate experiments investigated the effects of the length of the retention interval, the kind of processing during incidental learning, and the number of study versus test trials during intentional learning. The findings, together with a review of selected studies in the literature, support three conclusions: (1) Latency of correct recall is not necessarily more sensitive than accuracy, (2) accuracy and latency of correct recall measure different aspects of memory, and (3) latency of correct recall and latency of incorrect recall measure different aspects of memory. The available data disconfirm the idea that any unidimensional construct (such as strength or the amount of information in memory) underlies memory performance. An explanation is offered that emphasizes the distinction between encoding and retrieval processes.

A fundamental question about any dependent variable is: What does the variable measure? When two dependent variables co-exist a second question arises: Do these two variables measure the same thing? Two goals of the present article are (1) to explore the relationship between the two most frequently used indices of memory performance, namely,

* Order of authorship for this article is arbitrary. This research was supported by U.S. Public Health Service grants MH-21037 and MH-32205 and by Natural Sciences and Engineering Research Council Canada grant A7459. For their assistance in carrying out the experiments, we thank E. Ashbrook, P. Reinker, T. Ashbrook, and especially E. Witter.

Requests for reprints may be sent to Colin M. MacLeod, Division of Life Sciences, University of Toronto, Scarborough Campus, Scarborough, Ontario M1C 1A4 CANADA or to Thomas O. Nelson, Department of Psychology, University of Washington, Seattle, Washington 98195, USA.

response accuracy and response latency (both correct-response latency and error-response latency), and (2) to consider what aspects of memory each of these dependent variables might measure.

The relationship between response accuracy and response latency

The relationship between response accuracy and response latency has not been established, which may be one reason for the following conflicting beliefs. It is often assumed, as Pachella (1974) points out, "that the reaction time for correct responses is not affected by the overall error rate for an experimental condition. Such an assumption, in most instances, could not be more false ... *Average correct reaction time is inversely related to error rate*" (1974: 62, his italics). In contrast to this assumption of an inverse relationship, the opposite assumption has been made by other investigators (e.g., Baddeley and Ecob 1970; Hayes-Roth 1977).

One resolution to this apparent conflict concerns the different roles of response latencies. Pachella's statement refers to the well-known tradeoff between (a) the amount of time that the subject has available to respond before terminating the search through memory and (b) the likelihood that the target item will be retrieved. In general, the longer the subject searches through memory, the more likely the item is to be found. The amount of search time can be stipulated either by the experimenter (Wickelgren 1977) or by the subject (e.g., when further search seems unlikely to be successful). When the experimenter varies the amount of time allowed for memory search, the correct-response latency may approach the error-response latency (hereafter correct latency and error latency, respectively). Generally, such studies are designed so that subjects make few errors. The common practice is then to discard the error latencies from the analysis (Pachella et al. 1978), which is why Pachella's statement focused only on correct latency.

By contrast, when the subject determines the duration of search in a self-paced recall study, errors may be frequent. Concerning response latency, the typical pattern is that the average error latency is almost always longer than the average correct latency and, across conditions, there is a direct relation between correct latency and error rate (i.e., the speed-accuracy tradeoff). The two latency variables are analyzed separately and are interpreted in different ways. Correct latency is in-

terpreted as an index of the amount of information in memory about the item, and error latency is interpreted as an index of the subject's willingness to continue searching memory (Millward 1964).

These different interpretations of correct latency and error latency can be seen in the following examples. First, consider the view of correct latency as reflecting information in memory. The shorter the correct latency is, the more information there is presumed to be about that item in memory. This view has been upheld widely, but perhaps most clearly in studies of overlearning. Eimas and Zeaman (1963) showed that correct latency decreased as amount of overlearning increased. Note, however, that there is one precondition that seems necessary before this interpretation of correct latency can be adopted: Error latency must be longer than correct latency. Otherwise, any lack of a difference in correct latency across conditions could be an artifact of premature search termination, either experimenter-imposed or subject-imposed. Fortunately, error latency typically is of greater duration than correct latency in self-paced recall tasks (cf. Millward 1964).

Now consider error latency as a measure of criterion time for search termination. Thompson (1977) presented her subjects with a series of general-information questions and recorded their latencies when they were unable to recall the answer to a given question. Then she presented the subjects with each of the questions for which the correct answer had not been recalled, and the subject made a feeling-of-knowing judgment in terms of predicted likelihood of recognizing that item's answer. Finally, there was a forced-choice recognition test on the nonrecalled answers. The important result was that, across items for each subject, error latency was reliably correlated with feeling-of-knowing judgments but not with subsequent recognition performance. Thus, error latency reflects what the subject *believes* is in memory. This finding has been replicated and extended by Nelson et al. (1981).

Other interpretations of correct latency and error latency are possible, but the preceding discussion demonstrates that different mechanisms can reasonably be ascribed to these two latency measures.

Two assumptions about response accuracy and correct latency

Now consider two assumptions common in the memory literature related to accuracy, correct latency, and memory performance. Our goal

is to demonstrate that the picture is more complicated than these assumptions would suggest. This has important implications for an understanding of these dependent variables and ultimately for theories of memory.

The first assumption is that performance on a test of memory taps some *unidimensional* structure. This was a central assumption in strength theory (e.g., Hull 1943) and still appears frequently outside that theory. The upshot is that the two dependent variables should always change in a qualitatively similar fashion. As strength increases for a given set of items, the probability of an error on those items should decrease. Similarly, as strength increases, the correct latency on those items should decrease. Error probability and correct latency should be positively correlated. This has been stated explicitly in a number of theories, although it is often difficult to know how wide a domain was intended. Examples can be found in the theories of Cavanagh (1972), Suppes et al. (1966), and Norman and Wickelgren (1969). Although the idea of a multidimensional structure for memory has been proposed (e.g., Bower 1967; Underwood 1972), the unidimensionality assumption continues to be popular. For instance, Hayes-Roth makes the following core assumption in her theory: "The probability and speed of activation of an association by an appropriate stimulus are increasing functions of the strength of the association" (1977: 261). As another example, Ratcliff says that his retrieval theory "provides an intrinsic tie-up between reaction time and accuracy" (1978: 84), with both correct latency and error probability being fit by his single parameter *u* (for relatedness) – "all trace information is mapped onto a unidimensional variable, that of relatedness" (1978: 63).

The second assumption is a more specific version of the first. This assumption is that a major difference between the two dependent variables is their relative *sensitivity* for detecting information about items in memory. Always, the claim has been that correct latency is more sensitive than (or at least as sensitive as) error probability. The rationale for this claim is that correct latency may continue to show differences between two conditions even when accuracy is the same for both conditions. For example, Osgood writes, "Latency appears to be more *sensitive* than other measures because mean latencies continue to decrease with continued training long after both amplitude and frequency have reached stable values" (1953: 328, his italics). As another example, Shapiro writes, "While the typical measures of ...

number of errors do not reflect variations in FAS [free-association strength] ... this does not preclude the possibility that FAS still exerts some influence demonstrable by a more sensitive performance measure. One potentially more sensitive measure is response latency" (1968: 223–224). As a final example, Wearing and Montague write, "If trace strength is defined only by the correctness or incorrectness of an item ... A better way of assessing item strength is to use a measure that varies continuously and is not constrained by the definition of the criterion that is used. An appropriate measure is response latency, since it is generally regarded as a sensitive index of trace strength" (1970: 9). The greater sensitivity of correct latency than error probability appears to be widely assumed.

Whereas the unidimensionality assumption requires that the two dependent variables change in the same direction whenever a given experimental manipulation occurs, the sensitivity assumption goes one step further to assert that error probability should never detect a difference that is not detected by correct latency (although correct latency may detect a difference that is not detected by error probability). If cases can be observed in which two or more conditions differ in error probability but not in correct latency, then the sensitivity assumption is disconfirmed. Moreover, if cases can be observed in which two or more conditions differ in error probability in one direction and differ in correct latency in the opposite direction, then the unidimensionality assumption is disconfirmed.

These are the issues at which our three experiments are aimed. Based on our results and a review of the related literature, we will argue that both assumptions have now been disconfirmed empirically. We will also offer an alternative conception of these dependent variables.

Experiments

The goal of these experiments was to address the unidimensionality and sensitivity assumptions by investigating the relationship between response accuracy and response latency in three different kinds of situations. The variety of tasks was chosen specifically to permit greater generalizability of the conclusions regarding the accuracy-latency relationship. In experiment 1, the independent variable was the length of the long-term retention interval that followed intentional learning. In experiment 2, the independent variable was the kind of processing (semantic versus nonsemantic) during incidental learning. In experiment 3, the independent variable was the number of study trials versus test trials during intentional learning.

Experiment 1

Method

Subjects

The *S*s, tested individually, were 74 University of Washington undergraduates whose participation partially fulfilled a course requirement. Because three *S*s from the 3-week group and five from the 5-week group failed to recall any items correctly, their data were not included in the analyses. This left 66 *S*s: 24 in the 1-week group, 24 in the 3-week group, and 18 in the 5-week group.

Materials

The items, a subset of those used previously by Nelson (1971), were 20 number-noun pairs of the form 48-DOLLAR. The pairs were typed and made into slides. Also, four vowel–consonant pairs of the form E–K were prepared for the practice trials. Test slides were of the form 48 ? and E ?.

Apparatus

Stimuli were presented via a slide projector with tachistoscopic shutter. When this shutter opened to expose a test slide, a photo-electric cell activated a msec timer. When the *S* responded aloud into a throat microphone, a voice key stopped the timer. *S*s were told that their responses were being recorded, necessitating the microphone. However, they were not informed about the recording of response latencies until the experiment was completed.

Acquisition procedure

Prior to the first study-test trial, there were two practice trials using the four vowel–consonant pairs to familiarize the *S* with the procedure. Then the 20 number–noun pairs were shown individually at a 5-sec rate for silent study. Following study, *S*s counted digits for 30 sec to reduce recall from short-term memory. Then each number stimulus was presented for 5 sec and the *S* attempted to recall aloud the correct noun. The order of items at test was a different random order from that used at study. After the first study-test trial on all pairs, items correctly recalled were removed. Thus, the next study-test trial included only those items not acquired on the prior trial. This "drop-out" technique helped to minimize overlearning (Battig 1965) and was used over subsequent study-test trials until each item was correct exactly once, whereupon acquisition terminated. *S*s were dismissed without being informed of the delayed retention test to follow.

Retention test procedure

Either 1, 3, or 5 weeks after acquisition, the *S*s returned for the retention test. To minimize warm-up effects, one study-test trial of the four vowel–consonant pairs was administered first. This was followed by the self-paced retention test on the 20 number–noun pairs. The *S* was told to take as much time as needed to make a response. Although no time limit was imposed, some response was required on every

trial, even if the *S* eventually had to guess when unsure. The instructions emphasized accuracy, with no mention that response latency was being recorded. This procedure was used so that the *S* would not be tempted to adopt a fast-guess strategy that would increase error rate and perhaps obscure normal retrieval strategies. To encourage searching for the correct response, *S*s were informed that there would be only one test on each item. Each *S* had a different random test order, and no accuracy feedback was provided.

Results and discussion

All of the data in each of the three experiments have been analyzed using both nonparametric and parametric statistics. The conclusions hold uniformly, regardless of statistic, so only the more conventional parametric analyses are reported. In reporting our descriptive statistics, we present both the means and medians of individual *S* medians to demonstrate that the data are quite orderly, despite the relatively few observations per *S* per condition.

In experiment 1, an individual *S* contributed an error probability, a median correct latency, and a median error latency. Table 1 displays the results for both accuracy and latency. For the accuracy data shown at the top of the table, error probability differs reliably across the retention intervals, $F(2, 63) = 12.11$, $MSE = 0.03$, $p < 0.001$. As expected, error probability increases as the retention interval increases. For the error latency data shown at the bottom of the table, the difference across retention intervals is not significant, $F(2, 63) = 0.68$, $MSE = 0.68$, $p > 0.10$. However, error latencies are considerably longer than correct latencies at every retention interval. These two characteristics of the error latency data are consistent with the view of such data as measuring criterion time for search termination.

More important for present concerns is correct latency. Although in the direction expected, given the unidimensionality assumption and the results for error probability, correct latency does not differ reliably across the retention intervals, $F(2, 63) = 2.24$,

Table 1
Experiment 1: Mean and median error probability, correct latency, and error latency as a function of retention interval.

Dependent variable	Measure	Retention interval		
		1 week	3 weeks	5 weeks
Error probability	Mean	0.54	0.75	0.79
	Median	0.60	0.80	0.85
Correct latency	Mean	3.78	6.90	7.11
	Median	3.30	4.91	5.53
Error latency	Mean	18.12	17.91	14.97
	Median	14.66	16.24	14.74

$MSE = 34.22$, $p > 0.10$ [1]. Thus, even if the same pattern appears across conditions for correct latency as for error probability, correct latency can be less sensitive than error probability for discriminating between treatment groups. The next experiment converges on the same conclusion – that correct latency is not always as sensitive as error probability for discriminating between conditions.

Experiment 2

Method

Subjects

The *S*s, individually tested, were 27 University of Washington undergraduates whose participation partially fulfilled a course requirement. One additional *S* was discarded for making more than two errors during the classification task.

Materials and apparatus

The items were the 12 noun-noun paired associates shown in table 2. Each pair fell into one of the four cells of a 2×2 classification scheme. The two classification dimensions were (a) number of syllables (same vs different), and (b) size of referent (second member larger or smaller than first member of the pair). The items were prepared as slides for study, with the first word of each pair typed above the second

Table 2
Experiment 2: The noun-noun pairs.

	Structural classification (number of syllables)	
	Same number of syllables	Different number of syllables
Semantic classification (size of referents)		
First larger than second	FOREST-ELBOW	BARREL-COIN
	MULE-NAIL	VILLAGE-CHIN
	KETTLE-APPLE	HALL-INSECT
Second larger than first	GEM-DRESS	CLAW-TOWER
	FROG-CHAIR	FORK-OCEAN
	PEPPER-ARMY	BULLET-FLAG

[1] In fact, the apparent differences in correct latency over retention interval are due largely to a few subjects with particularly long median correct latencies. If those outliers greater than 8 sec are removed, then the mean correct latency for the 1-week group is 3.47 ($n = 23$), for the 3-week group is 3.68 ($n = 18$), and for the 5-week group is 3.52 ($n = 12$). Not surprisingly, these differences are nonsignificant, $F(2, 50) = 0.06$, $MSE = 3.17$, $p > 0.10$, and the sharp decrease in error variance from removing outliers strengthens the argument that correct latency is unaffected by retention interval.

word. At test, the upper word was presented as a cue for recall of the lower word. The apparatus was identical to that used in experiment 1.

Acquisition procedure

Acquisition of the list was incidental, with *S*s informed that they were performing a classification of the word pairs. The two questions used for the classification decision were: (a) SIZE – "Does the upper word of the pair represent something that is larger than the lower word?", and (b) SYLLABLES – "Does the upper word contain the same number of syllables as the lower word?" This structural-semantic distinction has been shown to have a large effect on incidental recall (Hyde and Jenkins 1969; Nelson 1977). Also, these particular questions necessitated that *S*s examine both of the words in a given pair before a YES or NO decision could be made.

There was one practice trial using a single non-list item for each type of decision to familiarize *S*s with the procedure. Then the first experimental trial began. Each *S* saw all 12 of the noun-noun pairs during a classification trial. Prior to presenting each pair, the experimenter said either "SIZE decision" or "SYLLABLE decision", indicating which classification to use on that trial. As soon as the pair appeared, the *S* responded "YES" or "NO" as rapidly as possible. *S*s were told that they were being timed, but were cautioned to avoid errors due to responding too quickly.

There were four successive classification trials on the entire list, each using a different item order. While they were not told that there would be four repetitions of each item, *S*s did know in advance that items would be repeated. On each trial: (a) No question appeared more than twice in succession, (b) there was an equal number of SYLLABLE and SIZE questions, and (c) there was an equal number of YES and NO answers, with never more than three of the same response in sequence.

Recall procedure

A blocking procedure was used to minimize recall from short-term memory (Nelson 1971). The fourth study trial had been divided into two 6-item blocks. Following this study trial, instructions for the incidental recall test were given. The recall trial was blocked in the same way as the fourth study trial, such that at least 6 items were studied and/or tested between the study and test of a given item. *S*s were told that they would see the upper member of each pair and that they should try to recall the lower member. A response was required on every trial, with *S*s guessing when necessary. Unlike experiment 1, *S*s were told that each response was being timed; however, they were encouraged not to sacrifice accuracy for speed.

Results and discussion

Classification trials

A 2×4 analysis of variance was conducted on the classification latencies displayed in table 3. Over the four trials, classification latencies decreased, $F(3, 66) = 29.65$, $MSE = 0.12$, $p < 0.001$. Although kind of processing had no overall effect on classification latency, $F(1, 22) = 0.31$, $MSE = 0.06$, $p > 0.10$, the marginally significant interaction, $F(3, 66) = 2.39$, $MSE = 0.04$, $0.10 > p > 0.05$, suggests that the decrease in

latency was slightly more marked for semantically processed items than for structurally processed items. Still, the main reason for presenting these data is to illustrate that kind of processing had little influence on classification latency.

Recall trial

Our main concern is with the results from the recall test; the means and medians are shown in table 4. In the latency analyses to be described, sample size varies somewhat due to some *S*s getting the items in a condition either all right or all wrong. When an *S* did not provide a correct latency or an error latency for both conditions, that *S* was excluded from the relevant analysis. Sample size is reflected by the corresponding degrees of freedom.

Error probability, shown at the top of the table, differs reliably across the two conditions, $t(26) = 4.51$, $MSE = 0.06$, $p < 0.001$. As has been reported widely, accuracy is greater after semantic processing than after structural processing. For the latency data, error latency is reliably longer than correct latency in the semantic condition, $t(17) = 2.57$, $MSE = 3.94$, $p < 0.05$, and in the structural condition, $t(21) = 3.20$, $MSE = 8.88$, $p < 0.01$. This pattern is consistent with that observed in experiment 1, and with the idea of error latencies as indexing willingness to continue search. An unanticipated finding was that recall error latency is reliably longer for items processed structurally than for items processed semantically, $t(15) = 3.37$, $MSE = 3.14$, $p < 0.01$. Apparently, *S*s are willing to search longer for structurally than for semantically processed items.

Table 3
Experiment 2: Mean classification latencies as a function of trial and kind of processing.

Kind of processing	Classification trial			
	1	2	3	4
Structural	1.72	1.46	1.25	1.23
Semantic	1.87	1.49	1.21	1.17

Table 4
Experiment 2: Mean and median error probability, correct latency, and error latency as a function of kind of processing.

Dependent variable	Measure	Kind of processing	
		Structural	Semantic
Error probability	Mean	0.56	0.29
	Median	0.60	0.20
Correct latency	Mean	4.44	2.66
	Median	2.35	2.30
Error latency	Mean	23.80	13.23
	Median	18.43	7.72

Turning to the correct latency data, there is no reliable difference during the recall of semantically processed items versus structurally processed items, $t(23) = 1.21$, $MSE = 1.47$, $p > 0.10$ [2]. Thus, as in experiment 1, the present pattern of results – in which error probability differs reliably across conditions while correct latency does not – again disconfirms the notion that correct latency is at least as sensitive as accuracy.

Taken together, the results of experiment 2 are consistent with the following account. The accuracy results suggest that structurally processed items are encoded and/or retrieved less effectively than are semantically processed items, when the memory test is paired-associate recall. The correct latency results imply that an item successfully encoded can be retrieved equally fast regardless of the kind of processing by which it was encoded. We will consider these ideas again following the next experiment.

Experiment 3

Method

Subjects, materials, and apparatus

The *S*s, tested individually, were 58 University of Washington undergraduates whose participation partially fulfilled a course requirement. The items were 16 of the 20 paired associates used in experiment 1. The practice items were three letter-letter pairs. The apparatus was identical to that in experiment 1.

Acquisition procedure

There were two groups of *S*s ($N = 29$ per group), a multiple-study group (SSST) and a multiple-test group (STTT). *S*s in both groups had a single study-test trial on the three practice pairs before beginning the actual experiment. Then all *S*s were told that there would be a variable number of study and test trials on the 16-pair list, but neither group was told the precise number of study or test trials in advance. For Group SSST, acquisition consisted of three consecutive study trials followed by one test trial. For Group STTT, acquisition consisted of a single study trial and three successive test trials, with a different testing order on each test trial. A maximum of 8 sec was allowed for the test of each item, with latencies recorded for responses faster than 8 sec. There was a 2-sec inter-item interval. *S*s were informed that their responses were being timed, but were cautioned to emphasize accuracy over speed.

Retention procedure

Following a 5-min period of solving a block puzzle (to minimize recall from

[2] As the medians suggest, the apparent difference in overall mean correct latency (which is in the opposite direction from that predicted by the unidimensional assumption) is due largely to a few subjects with particularly long median correct latencies. If the four subjects with outliers beyond 8 sec are removed, then the means become 2.00 for the structural condition and 2.70 for the semantic condition, again a nonsignificant difference, $t(19) = 1.53$, $MSE = 0.46$, $p > 0.10$. Notice in particular the reduction in error variance from removing these few outliers.

short-term memory), a self-paced retention test of the 16 items was administered. For each item, the *S* attempted to say the noun that went with the number stimulus. A response was required for every item, with the *S* guessing when necessary. The *S* was aware that latencies were being collected, and the test order was a new randomization of all 16 numbers.

Results and discussion

Acquisition trials

The results for error probability, correct latency, and error latency on each test trial during acquisition are shown in table 5. For the STTT group, the small differences in error probability across the three test trials were marginally reliable, $F(2, 56) = 2.99$, $MSE = 0.004$, $0.10 > p > 0.05$, although the pattern has no obvious theoretical interpretation. While the decrease in correct latency across the three test trials was substantial, it did not reach conventional significance levels, $F(2, 54) = 2.25$, $MSE = 0.96$, $p > 0.10$. Because error latencies had an imposed upper boundary of 8 sec during acquisition (and table 5 shows that the means were generally quite close to that boundary), we will not discuss error latency in detail. It is worth noting, however, that error latencies were significantly longer than their corresponding correct latencies in every case.

Of primary importance are the comparisons of the STTT group with the SSST group on the final trial of the acquisition phase and on the retention test trial. As table 5 indicates, error probability was reliably lower on the final acquisition trial for the SSST group than for the STTT group, $t(56) = 5.94$, $MSE = 0.06$, $p < 0.001$; by contrast, correct latency was higher for the SSST group than for the STTT group, although this difference was not reliable, $t(56) = 0.74$, $MSE = 0.30$, $p > 0.10$.

Retention test trial

As comparison of tables 5 and 6 demonstrates, essentially the same pattern of results occurred on the retention test as on the final acquisition test. Error probability was again reliably lower for the SSST group than for the STTT group, $t(56) = 5.98$, $MSE = 0.06$, $p < 0.001$, whereas correct latency was again higher for the SSST group

Table 5
Experiment 3: Mean and median error probability, correct latency, and error latency on acquisition test trials for the STTT and SSST groups.

Dependent variable	Measure	STTT acquisition			SSST acquisition
		Test 1	Test 2	Test 3	Test
Error probability	Mean	0.70	0.73	0.69	0.35
	Median	0.75	0.75	0.69	0.38
Correct latency	Mean	2.47	2.36	1.94	2.26
	Median	2.32	1.96	1.96	1.63
Error latency	Mean	6.97	7.04	6.50	7.03
	Median	7.84	8.00	8.00	7.93

Table 6
Experiment 3: Mean and median error probability, correct latency, and error latency on the retention test trial for the STTT and SSST groups.

Dependent variable	Measure	Group	
		STTT	SSST
Error probability	Mean	0.65	0.29
	Median	0.69	0.25
Correct latency	Mean	1.95	2.05
	Median	1.57	1.83
Error latency	Mean	29.87	19.32
	Median	20.48	12.82

than for the STTT group, although the difference was not statistically reliable, $t(56) = 0.46$, $MSE = 0.23$, $p > 0.10$. Error latency was free to vary on the retention test, but the difference between conditions was not significant, $t(56) = 1.34$, $MSE = 7.84$, $p > 0.10$.

Thus, as in experiments 1 and 2, the present pattern of results – in which error probability differs reliably across conditions while correct latency does not – disconfirms the notion that latency is at least as sensitive as accuracy. Moreover, although each of the two differences in correct latency between the SSST group and the STTT group fell short of statistical significance, both differences in correct latency were in the opposite direction from the corresponding differences in error probability. Such reversals are in violation of the assumption that latency and accuracy measure the same dimension in tests of memory.

How might one interpret these data? The scheme outlined at the end of experiment 2 seems useful. The STTT group must rely on only a single encoding opportunity, resulting in steady-state accuracy over test trials. Retrieval practice decreases correct latency as *S*s become familiar with access routes. By contrast, the SSST group systematically builds up efficient encodings of the items over successive study trials, resulting in higher accuracy on the final test trial. However, without the opportunity to practice retrieval, correct latencies are long. These ideas are consistent with those of previous researchers (Birnbaum and Eichner 1971; Hogan and Kintsch 1971) concerning the notion that relative to test trials, study trials have a larger effect on encoding but a smaller effect on retrieval. Further, the present results suggest that the dependent variables of accuracy and correct latency are differentially sensitive to these encoding and retrieval effects.

General discussion

The relative sensitivities of latency and accuracy

In all three experiments, the results disconfirmed the prevalent assumption that correct latency is at least as sensitive as error probability in

differentiating between two or more conditions [3]. The conditions under investigation were reliably different when assessed by error probability but not when assessed by correct latency. This occurred for conditions that differed in terms of length of retention interval (experiment 1), kind of processing during incidental learning (experiment 2), and study versus test trials during intentional learning (experiment 3).

This is not the only disconfirmation of the sensitivity assumption. Scheirer (1971) found reliable differences in error probability but not in correct latency for three independent variables: (1) modality of presentation (auditory vs visual), (2) duration of presentation (1 vs 4 sec), and (3) direction of recall (forward vs backward). As well, unpublished research in our laboratory showed that overlearning during acquisition subsequently had a reliable effect on error probability but not on correct latency in long-term retention. Thus, the domain of disconfirmation for the assumption that correct latency is more sensitive than accuracy is fairly broad. Indeed, the findings are more in accord with the hypothesis that response accuracy is at least as sensitive as (or more sensitive than) correct latency! How limited is the domain of this hypothesis?

At first glance, disconfirmation might seem to be ubiquitous. For

[3] Two points should be made about the conceptualization of sensitivity. First, the critical data concerning accuracy and correct latency come from the same number of items, namely, the number of items correct for a given subject in a given condition. This may not be obvious in the case of accuracy because the reported accuracy score is the error probability (i.e., "number of errors" divided by "total number of items"). Alternatively, the accuracy score could have been the number correct, obtained via linear transformation of error probability (i.e., "number correct" = "total number of items" minus the product of "error probability" and "total number of items"), without affecting the F or t value of any statistical comparison. Second, as in Nelson (1977), sensitivity refers to one test detecting a difference between two conditions that another test does not detect. Ideally, this would refer to population values; in empirical research, however, the comparison is on sample values, which are not perfectly reliable as estimates of their respective population values. Hence, the issues of validity and reliability are intertwined when the relative sensitivities of two tests are compared in empirical situations. Moreover, there typically is no way of mapping the units of one test (e.g., sec) and the units of the other test (e.g., number of correct responses) into the same scale so that they can be compared directly. Nevertheless, the present results are sufficient to demonstrate that correct latency is sometimes less sensitive than accuracy for the statistically reliable detection of differences that we manipulated via the independent variable. This conceptualization of sensitivity seems necessary whenever the measures are sample estimates of population values (rather than population values per se) and is consistent with previous usage in the literature cited above.

instance, almost every study using the Sternberg (1966) paradigm seems at odds with this idea. However, those findings are inappropriate for a theoretically relevant test of the hypothesis because they contain a ceiling effect on observed accuracy, making it impossible to find that accuracy is more sensitive than correct latency. The problem is endemic. We could not find even one report in any area of memory research where a reliable difference occurred for correct latency but not for error probability *and* in which a ceiling effect on accuracy was not present. Thus, when performance on neither measure is at the floor or ceiling, there is no known exception to the hypothesis that response accuracy is at least as sensitive as correct latency.

This observation highlights an important methodological issue. When a ceiling effect on accuracy is likely, then obtaining latencies is desirable if a difference between conditions is to be detected. This might be important in an applied situation, for example. However, if the goal is to draw theoretical conclusions, then the ceiling effect on accuracy disallows theoretical conclusions about (the lack of) differential effects on accuracy. Moreover, the ceiling effect on accuracy also produces problems for theoretical conclusions about differential effects on latency. In their review of latency research, Pachella et al. make this point forcefully:

> If subjects actually produced no errors in an experiment, the theorist would be at a loss to interpret the obtained reaction times, because there are an infinite number of reaction times that can result in zero errors ... small differences in error rate can be associated with large differences in reaction time. This is particularly true for the range of high overall accuracy (90% to 100%) typically found in reaction-time experiments. This means that what may look like relatively meaningless differences in accuracy may contaminate reaction-time values extensively. (1978: 172)

This point, perhaps not recognized a decade ago, has also been made by Pachella (1974) and Wickelgren (1977).

Because there currently are no disconfirmations of the hypothesis that accuracy is at least as sensitive as correct latency, whereas there are disconfirmations of the suggestion that correct latency is at least as sensitive as accuracy, one might be tempted to conclude that accuracy is strictly more sensitive than correct latency (except when a ceiling or floor effect on accuracy is present). Unfortunately, the problem is more complex than this, largely because of the issue of the dimensionality of memory.

Unidimensional versus multidimensional memory structure

Underlying the question of relative sensitivity is the notion of a unidimensional structure for memory. If memory were unidimensional, then the difference in error probability for any given pair of conditions would always be in the same direction as the difference in correct latency for that pair of conditions. An empirical outcome of opposite directions of difference for error probability and correct latency would be sufficient to disconfirm the assumption that the underlying memory structure is unidimensional, thereby requiring that it be multidimensional. Consequently, the crucial issue is whether any reliable disconfirming empirical outcome has occurred regarding unidimensionality.

The results of experiment 3 hinted at such an outcome. Error probability was reliably lower for SSST than for STTT, whereas correct latency was higher for SSST than for STTT. However, because the difference in correct latency was not reliable on either test trial, these results can be taken as suggestive only. Meanwhile, other studies in the memory literature do provide reliable evidence for opposite directions of difference for error probability and correct latency.

Peterson et al. (1977) investigated the effects on recall of imagery versus rote learning instructions. In forward recall, the values of error probability were reliably lower whereas the values of correct latency were reliably higher following imaginary versus rote learning. The same pattern occurred reliably for backward recall. In a different situation, Corbett (1977) used a speed-accuracy tradeoff method to study recognition and found that asymptotic accuracy was higher for imagery than for rote learning, whereas retrieval was slower for imagery than for rote learning.

A similar pattern occurs in studies comparing verbal mediation and rote learning. For instance, Wearing and Montague (1970) found a situation in which the values of error probability were lower whereas the values of correct latency were higher for verbal mediation than for rote learning. As Adams says, after explaining that verbal mediators reduce error probability, "An NLM [natural language mediator] only makes the response to the pair longer" (1967: 305). Thus, mediated learning – whether imaginal or verbal – seems to increase accuracy but decrease speed, which disconfirms the assumption of a unidimensional structure underlying memory performance. Different processes can affect these two dependent variables in qualitatively different ways.

Other evidence also supports the conclusion that the underlying memory structure is multidimensional [4]. In a study by Arbak (reported in Murdock 1974: 101–102), each paired associate received one study trial and two recall test trials. The hypothesis derived from the notion of a unidimensional structure underlying memory performance was that pairs with a shorter correct latency on the first test trial should have a lower error probability on the second test trial. One experiment yielded no relationship, while the other experiment actually yielded the opposite relationship, namely, items having a shorter correct latency on the first test trial showed a higher error probability on the second test trial (also see Jacoby 1978).

As a final example, one can explore an individual differences analysis, in accord with Underwood's (1975) suggestion. Scheirer (1971) examined the correlation between error probability and correct latency across 89 subjects. He reasoned that if error probability and correct latency were measuring the same underlying factor, then a high positive correlation should occur. Instead, however, the correlation was $r = 0.18$, which is not reliably different from zero. Lachman and Lachman (1980) reach the same conclusion; correct latency and error probability are uncorrelated.

Thus, contrary to views expressed by many previous researchers, accuracy and correct latency do not seem to be indices of a single underlying dimension of memory structure. An analogy can be made to the physical sciences where height and weight are indices of different dimensions of physical objects. Although somewhat correlated, height and weight are not interchangeable for qualitatively ordering various objects. The same seems to be the case with error probability and correct latency. This lack of interchangeability may help to explain why some researchers have had difficulty reconciling their findings with previous findings. For instance, in a priming study, Brown (1979) found

[4] Not all evidence adduced in support of multidimensionality is equally convincing. For instance, Anderson and Bower (1972) focused on the opposite direction of the word-frequency effect in recall versus recognition. However, such a comparison implicitly assumes a theory of each task, and the validity of the overall conclusion depends upon the validity of those theories (otherwise, any differences could be due to artifacts of the tasks). In particular, to claim that recognition is more accurate for low-frequency words than for high-frequency words requires a theory of recognition that encompasses the distractors, and the version espoused by Anderson and Bower is currently controversial (Hall 1983). An advantage of the present research is that the comparison is not between two tasks, but instead is between two measures obtained from a single task.

results opposite to those of three previously reported studies; he used latency as the dependent variable in his study while all three of the cited studies with the opposite conclusion had used accuracy.

To disconfirm unidimensionality is to disconfirm any underlying structure consisting of one fundamental dimension utilizing numerosity (e.g., *amount* or *number* of something, such as amount of memory strength or amount of information). Theoreticians who choose to retain the core assumption of unidimensionality will have to invoke auxiliary assumptions to account for the opposite effects of some independent variables on accuracy versus correct latency. The other possibility is for theoreticians to postulate an underlying structure that is multidimensional.

One reason that multidimensionality may seem counterintuitive for speed and accuracy is that fast responses often tend to be correct while slow responses tend to be error-prone. Yet there are also counterexamples. A common example occurs in answering the question, "How many days are there in November?" People are accurate but slow as they proceed through the rhyme "thirty days hath September, April, June, and November ..." Examples also occur in which responses are fast but inaccurate. For instance, when someone has recently moved and has a new telephone number, they might respond quickly but inaccurately when asked for their current telephone number or, when someone has recently married and changed their name, they might respond quickly but inaccurately when asked for their current name. The reader probably can generate other examples that vary differently across error probability and correct latency.

An alternative conception

These three new experiments and a brief review of the relevant literature call into question some common assumptions about the fundamental postulates that relate response latency and response accuracy to memory performance and memory structure. It no longer seems reasonable to assume that correct latency is more sensitive than response accuracy, and it seems inappropriately simplistic to conceptualize the underlying memory structure as unidimensional. These issues are important because they influence both the design of experiments on memory and the explanation of the outcomes of such experiments.

The conception offered here construes error probability as a measure of the likelihood that the encoding of an item was sufficient for it to be retrieved when given the cues of the test environment. By contrast, correct latency is construed as a measure of the number of decoding steps during retrieval before the item is output (e.g., the number of transformations while decoding a mediator, in Prytulak 1971). Error latency is construed as an index of the degree of willingness to continue search for a currently unretrieved item. Similar ideas have been put forth by Anderson (1981) in his analysis of interference effects in paired-associate learning, although his position on the sensitivity issue differs from ours. The key point is that these dependent variables are not measuring identical processes, although they may overlap on some occasion (i.e., the occurrence of the same qualitative effect on accuracy and correct latency is not inconsistent with the concept of a multidimensional structure).

Conclusions

(1) Response latencies should be segregated in terms of correct latency and error latency. Earlier studies that combined them usually found that the combined latency increased as error probability increased (e.g., Shapiro 1968). However, such a finding might be due to increases in error latency alone or to the fact that error latencies typically are longer than correct latencies (Millward 1964), which the combined latency will reflect. Moreover, error latency and correct latency should be interpreted as reflecting different facets of memory, with error latency construed as an index of willingness to continue searching, and correct latency construed as a measure of retrieval from memory.

(2) New findings reported here, along with those from previous studies, disconfirm the hypothesis that correct latency is more sensitive than error probability as an index of memory. Indeed, the present findings demonstrated that error probability was more sensitive than correct latency for reliably detecting the difference between two conditions. Moreover, no evidence was found in the literature for the hypothesis that correct latency is more sensitive than error probability *in the absence of floor and ceiling effects.*

(3) Error probability and correct latency seem to tap different dimensions of the memory structure. One possibility is that error probability measures the sufficiency of the encoding for retrieval, whereas correct

latency measures the number of decoding steps during retrieval before the item is output. Hence, neither correct latency nor error probability is inherently preferable. When only one of these measures is to be studied, the choice should be determined by which facets of memory are of interest.

References

Adams, J.A., 1967. Human memory. New York: McGraw-Hill.

Anderson, J.R., 1981. Interference: the relationship between response latency and response accuracy. Journal of Experimental Psychology: Human Learning and Memory 7, 326–343.

Anderson, J.R. and G.H. Bower, 1972. Recognition and retrieval processes in free recall. Psychological Review 79, 97–123.

Atkinson, R.C. and J.F. Juola, 1974. 'Search and decision processes in recognition memory'. In: D.H. Krantz, R.C. Atkinson, R.D. Luce and P. Suppes (eds.), Contemporary developments in mathematical psychology, Vol. 1. San Francisco, CA: Freeman.

Baddeley, A.D. and J.R. Ecob, 1970. Reaction time and short-term memory: a trace strength alternative to the high-speed exhaustive scanning hypothesis. Technical report number 13, Center for Human Information Processing, University of California, San Diego.

Battig, W.F., 1965. Procedural problems in paired-associate learning research. Psychonomic Monograph Supplements 1.

Birnbaum, I.M. and J.T. Eichner, 1971. Study versus test trials and long-term retention in free-recall learning. Journal of Verbal Learning and Verbal Behavior 10, 516–521.

Bower, G.H., 1967. 'A multicomponent theory of the memory trace'. In: K.W. Spence and J.T. Spence (eds.), The psychology of learning and motivation, Vol. 1. New York: Academic Press.

Brown, A.S., 1979. Priming effects in semantic memory retrieval processes. Journal of Experimental Psychology: Human Learning and Memory 5, 65–77.

Cavanagh, J.P., 1972. Relation between the immediate memory span and the memory search rate. Psychological Review 79, 525–530.

Corbett, A.T., 1977. Retrieval dynamics for rote and visual image mnemonics. Journal of Verbal Learning and Verbal Behavior 16, 233–246.

Eimas, P.D. and D. Zeaman, 1963. Response speed changes in an Estes' paired-associate 'miniature' experiment. Journal of Verbal Learning and Verbal Behavior 1, 384–388.

Hall, J.F., 1983. Recall versus recognition: a methodological note. Journal of Experimental Psychology: Learning, Memory, and Cognition 9, 346–349.

Hayes-Roth, B., 1977. Evolution of cognitive structures and processes. Psychological Review 84, 260–278.

Hogan, R.M. and W. Kintsch, 1971. Differential effects of study and test trials on long-term recognition and recall. Journal of Verbal Learning and Verbal Behavior 10, 562–567.

Hull, C.L., 1943. Principles of behavior. New York: Appleton-Century-Crofts.

Hyde, T.S. and J.J. Jenkins, 1969. Differential effects of incidental tasks on the organization of recall of a list of highly associated words. Journal of Experimental Psychology 82, 472–481.

Jacoby, L.L., 1978. On interpreting the effects of repetition: solving a problem versus remembering a solution. Journal of Verbal Learning and Verbal Behavior 17, 649–667.

Lachman, J.L. and R. Lachman, 1980. 'Age and actualization of word knowledge'. In: L.W. Poon, J.L. Fozard, L.S. Cermak, D. Arenberg and L.W. Thompson (eds.), New directions in memory and aging. Hillsdale, NJ: Erlbaum.

Millward, R., 1964. Latency in a modified paired associate learning experiment. Journal of Verbal Learning and Verbal Behavior 3, 309–316.

Murdock, B.B. Jr., 1974. Human memory: theory and data. New York: Wiley.

Nelson, T.O., 1971. Savings and forgetting from long-term memory. Journal of Verbal Learning and Verbal Behavior 10, 568–576.

Nelson, T.O., 1977. Repetition and depth of processing. Journal of Verbal Learning and Verbal Behavior 16, 151–171.

Nelson, T.O., R.F. Landwehr, R.J. Leonesio, J.L. Raley, A.M. Weisman and L. Narens, 1981. Predictive validity of feeling-of-knowing judgments for the recognition of nonrecalled general-information facts. Paper presented at the annual meeting of the Psychonomic Society, Philadelphia, PA.

Norman, D.A. and W.A. Wickelgren, 1969. Strength theory of decision rules and latency in short-term memory. Journal of Mathematical Psychology 6, 192–208.

Osgood, C.E., 1953. Method and theory in experimental psychology. New York: Oxford University Press.

Pachella, R.G., 1974. 'The interpretation of reaction time in information processing research'. In: B.H. Kantowitz (ed.), Human information processing: tutorials in performance and cognition. New York: Wiley.

Pachella, R.G., J.E.K. Smith and K.E. Stanovich, 1978. 'Qualitative error analysis and speeded classification'. In: N.J. Castellan and F. Restle (eds.), Cognitive theory, Vol. 3. Hillsdale, NJ: Erlbaum.

Perlmutter, J., P. Sorce and J.L. Myers, 1976. Retrieval processes in recall. Cognitive Psychology 8, 32–63.

Peterson, L.R. and M.J. Peterson, 1959. Short-term retention of individual verbal items. Journal of Experimental Psychology 58, 193–198.

Peterson, L.R., L. Rawlings and C. Cohen, 1977. 'The internal construction of spatial patterns'. In: G.H. Bower (ed.), The psychology of learning and motivation, Vol. 11. New York: Academic Press.

Prytulak, L.S., 1971. Natural language mediation. Cognitive Psychology 2, 1–56.

Ratcliff, R., 1978. A theory of memory retrieval. Psychological Review 85, 59–108.

Scheirer, C.J., 1971. Effect of cueing, modality, and effective contiguous time on response latency in short-term memory. Journal of Experimental Psychology 88, 429–432.

Shapiro, S.I., 1968. Paired-associate response latencies as a function of free association strength. Journal of Experimental Psychology 77, 223–231.

Shepard, R.N., 1967. Recognition memory for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior 6, 156–163.

Sternberg, S., 1966. High-speed scanning in human memory. Science 153, 652–654.

Suppes, P., G. Groen and M. Schlag-Rey, 1966. A model for response latency in paired-associate learning. Journal of Mathematical Psychology 3, 99–128.

Thompson, B.G., 1977. The feeling of knowing: decision to terminate the search. Master's thesis from the University of Houston.

Tulving, E. and G.H. Bower, 1974. 'The logic of memory representations'. In: G.H. Bower (ed.), The psychology of learning and motivation, Vol. 8. New York: Academic Press.

Tulving, E. and Z. Pearlstone, 1966. Availability versus accessibility of information in memory for words. Journal of Verbal Learning and Verbal Behavior 5, 381–391.

Underwood, B.J., 1972. 'Are we overloading memory?' In: A.W. Melton and E. Martin (eds.), Coding processes in human memory. New York: Wiley.

Underwood, B.J., 1975. Individual differences as a crucible in theory construction. American Psychologist 30, 128–134.

Wearing, A.J. and W.F. Montague, 1970. A test of the Battig procedure for controlling the level of individual item learning in paired-associate lists. Behavioral Research Methods and Instrumentation 2, 9–10.

Wescourt, K.T. and R.C. Atkinson, 1976. 'Fact retrieval processes in human memory'. In: W.K. Estes (ed.), Handbook of learning and cognitive processes, Vol. 4. Hillsdale, NJ: Erlbaum.

Wickelgren, W.A., 1977. Speed-accuracy tradeoff and information processing dynamics. Acta Psychologica 41, 67–85.