JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 8, 240–247 (1969)

Retrieval Time from Semantic Memory[1]

ALLAN M. COLLINS AND M. ROSS QUILLIAN

Bolt Beranek and Newman, Inc., Cambridge, Massachusetts 02138

To ascertain the truth of a sentence such as "A canary can fly," people utilize long-term memory. Consider two possible organizations of this memory. First, people might store with each kind of bird that flies (e.g., canary) the fact that it can fly. Then they could retrieve this fact directly to decide the sentence is true. An alternative organization would be to store only the generalization that *birds* can fly, and to infer that "A canary can fly" from the stored information that a canary is a bird and birds can fly. The latter organization is much more economical in terms of storage space but should require longer retrieval times when such inferences are necessary. The results of a true-false reaction-time task were found to support the latter hypothesis about memory organization.

Quillian (1967, 1969) has proposed a model for storing semantic information in a computer memory. In this model each word has stored with it a configuration of pointers to other words in the memory; this configuration represents the word's meaning. Figure 1 illustrates the organization of such a memory structure. If what is stored with canary is "a yellow bird that can sing" then there is a pointer to bird, which is the category name or *superset* of canary, and pointers to two *properties*, that a canary is yellow and that it can sing. Information true of birds in general (such as that they can fly, and that they have wings and feathers) need not be stored with the memory node for each separate kind of bird. Instead, the fact that a canary can fly can be inferred by retrieving that a canary is a bird and that birds can fly. Since an ostrich cannot fly, we assume this information is stored as a property with the node for ostrich, just as is done in a dictionary, to preclude the inference that an ostrich can fly. By organizing the memory in this way, the amount of space needed for storage is minimized.

If we take this as a model for the structure of human memory, it can lead to testable predictions about retrieving information. Suppose a person has only the information shown in Fig. 1 stored on each of the nodes. Then to decide "A canary can sing," the person need only start at the node canary and retrieve the properties stored there to find the statement is true. But, to decide that "A canary can fly," the person must move up one level to bird before he can retrieve the property about flying. Therefore, the person should require more *time* to decide that "A canary can fly" than he does to decide that "A canary can sing." Similarly, the person should require still longer to decide that "A canary has skin," since this fact is stored with his node for animal, which is yet another step removed from canary. More directly, sentences which themselves assert something about a node's supersets, such as "A canary is a bird," or "A canary is an animal," should also require decision times that vary directly with the number of levels separating the memory nodes they talk about.

A number of assumptions about the

[1] This research was supported by the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, under Contract No. F33615-67-C-1982 with Bolt Beranek and Newman, Inc. and also partly by Advanced Research Projects Agency, monitored by the Air Force Cambridge Research Laboratories, under Contract No. F19628-68-C-0125.

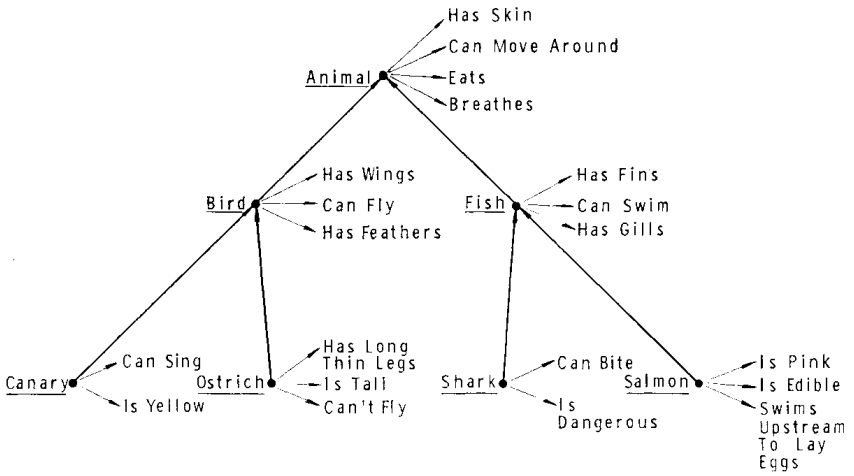

FIG. 1. Illustration of the hypothetical memory structure for a 3-level hierarchy.

retrieval process must be made before predictions such as those above can be stated explicitly. First, we need to assume that both retrieving a property from a node and moving up a level in a hierarchy take a person time. Second, we shall assume that the times for these two processes are additive, whenever one step is dependent on completion of another step. This assumption is equivalent to Donders' assumption of additivity (Smith, 1968) for the following two cases: (a) When moving up a level is followed by moving up another level, and (b) when moving up a level is followed by retrieving a property at the higher level. Third, we assume that the time to retrieve a property from a node is independent of the level of the node, although different properties may take different times to retrieve from the same node. It also seems reasonable to assume that searching properties at a node and moving up to the next level occur in a parallel rather than a serial manner, and hence are not additive. However, this assumption is not essential, and our reasons for preferring it are made clear in the Discussion section.

We have labeled sentences that state property relations P sentences, and those that state superset relations S sentences. To these labels numbers are appended. These indicate the number of levels the model predicts it would be necessary to move through to decide the sentence is true. Thus, "A canary can sing" would be a PO sentence, "A canary can fly" would be a P1 sentence, and "A canary has skin" would be a P2 sentence. Similarly, "A canary is a canary" would be an SO sentence, "A canary is a bird" would be an S1 sentence, and "A canary is an animal" would be an S2 sentence.

It follows from the assumptions above that the time differences predicted for PO, P1, and P2 sentences are entirely a result of moving from one level in the hierarchy to the next. Thus, the increase in time from SO to S1 should be the same as from PO to P1 since both increases are a result of moving from level O to level 1. Likewise, the time increase from S1 to S2 should equal the time increase from P1 to P2. In fact, if we assume that the time to move from one level to the next is not dependent on which levels are involved, all the time increases (from PO to P1, P1 to P2, SO to S1, and S1 to S2) should be equal.

Recently, reaction time (RT) has been used as a measure of the time it takes people to retrieve information from memory. By constructing a large number of true sentences of the six types discussed and interspersing these with equal numbers of false sentences, we can measure the reaction time for *S*s to decide which sentences are true and which are false. Thus, this method can be used to test the

prediction we have derived from the model and our assumptions about the retrieval process.

A caution is in order here: Dictionary definitions are not very orderly and we doubt that human memory, which is far richer, is even as orderly as a dictionary. One difficulty is that hierarchies are not always clearly ordered, as exemplified by dog, mammal, and animal. Subjects tend to categorize a dog as an animal, even though a stricter classification would interpose the category mammal between the two. A second difficulty is that people surely store certain properties at more than one level in the hierarchy. For example, having leaves is a general property of trees, but many people must have information stored about the maple leaf directly with maple, because of the distinctiveness of its leaf. In selecting examples, such hierarchies and instances were avoided. However, there will always be *S*s for whom extensive familiarity will lead to the storing of many more properties (and sometimes supersets) than we have assumed. By averaging over different examples and different subjects, the effect of such individual idiosyncrasies of memory can be minimized.

Method

Three experiments were run, with eight *S*s used in each experiment. The *S*s were all employees of Bolt Beranek and Newman, Inc. who served voluntarily and had no knowledge of the nature of the experiment. Because of a faulty electrical connection, only three *S*s gave usable data in Expt. 3. The same general method was used for all three experiments, except in the way the false sentences were constructed.

Apparatus. The sentences were displayed one at a time on the cathode ray tube (CRT) of a DEC PDP-1 computer.[2] The timing and recording of responses were under program control.[3] Each sentence was centered vertically on one line. The length of line varied from 10 to 34 characters (approximately 4–11° visual angle). The *S* sat directly in front of the CRT with his two index fingers resting on the two response buttons. These each required a displacement of $\frac{1}{4}$ in to trigger a microswitch.

Procedure. The sentences were grouped in runs of 32 or 48, with a rest period of approximately 1 min between runs. Each sentence appeared on the CRT for 2 sec, and was followed by a blank screen for 2 sec before the next sentence. The *S* was instructed to press one button if the sentence was generally true, and the other button if it was generally false, and he was told to do so as accurately and as quickly as possible. The *S* could respond anytime within the 4 sec between sentences, but his response did not alter the timing of the sentences. Each *S* was given a practice run of 32 sentences similarly constructed.

Sentences. There were two kinds of semantic hierarchies used in constructing sentences for the experiments, 2-level and 3-level. In Fig. 1, a 2-level hierarchy might include bird, canary, and ostrich and their properties, whereas the whole diagram represents a 3-level hierarchy. A 2-level hierarchy included true PO, P1, SO and S1 sentences; a 3-level hierarchy included true P2 and S2 sentences as well. Examples of sentence sets with 2-level and 3-level hierarchies are given in Table 1.[4] As illustrated in Table 1, equal numbers of true and false sentences were always present (but in random sequence) in the sentences an *S* read. Among both true and false sentences, there are the two general kinds: Property relations (P), and superset relations (S).

In Expt 1, each *S* read 128 two-level sentences followed by 96 three-level sentences. In Expt 2, each *S* read 128 two-level sentences, but different sentences from those used in Expt 1. In Expt 3, a different group of *S*s read the same 96 three-level sentences used in Expt 1. Each run consisted of sentences from only four subject-matter hierarchies.

To generate the sentences we first picked a hierarchical group with a large set of what we shall call *instances* at the lowest level. For example, baseball, badminton, etc. are instances of the superset game. Different instances were used in each sentence, because repetition of a word is known to have substantial effects in reducing RT (Smith, 1967). In constructing S1 and S2 sentences, the choice of the category name or superset was in most cases obvious, though in a case such as the above 2-level example, sport might have been used as the superset rather than game. To assess how well our choices corresponded

[2] Now at the University of Massachusetts, Amherst.

[3] The authors thank Ray Nickerson for the use of his program and for his help in modifying it to run on BBN's PDP-1.

[4] To obtain the entire set of true sentences for Expt 1 order NAPS Document NAPS-00265 from ASIS National Auxiliary Publications Service, c/o CCM Information Sciences, Inc., 22 West 34th Street, New York, New York 10001; remitting $1.00 for microfiche or $3.00 for photocopies.

TABLE 1
ILLUSTRATIVE SETS OF STIMULUS SENTENCES

	Sentence type	True sentences	Sentence type[a]	False sentences
Expt 1, 2-level				
	PO	Baseball has innings	P	Checkers has pawns
	P1	Badminton has rules	P	Ping pong has baskets
	SO	Chess is chess	S	Hockey is a race
	S1	Tennis is a game	S	Football is a lottery
Expt 1, 3-level				
	PO	An oak has acorns	P	A hemlock has buckeyes
	P1	A spruce has branches	P	A poplar has thorns
	P2	A birch has seeds	P	A dogwood is lazy
	SO	A maple is a maple	S	A pine is barley
	S1	A cedar is a tree	S	A juniper is grain
	S2	An elm is a plant	S	A willow is grass
Expt 2, 2-level				
	PO	Seven-up is colorless	PO	Coca-cola is blue
	P1	Ginger ale is carbonated	P1	Lemonade is alcoholic
	SO	Pepsi-cola is Pepsi-cola	SO	Bitter lemon is orangeade
	S1	Root beer is a soft drink	S1	Club soda is wine

[a]There were no distinctions as to level made for false sentences in Expt 1.

with the way most people categorize, two individuals who did not serve in any of the three experiments were asked to generate a category name for each S1 and S2 sentence we used, e.g., "tennis is ————." These two individuals generated the category names we used in about 3/4 of their choices, and only in one case, "wine is a drink" instead of "liquid", was their choice clearly not synonymous.

In generating sentences that specified properties, only the verbs "is," "has," and "can" were used, where "is" was always followed by an adjective, "has" by a noun, and "can" by a verb. To produce the PO sentence one of the instances such as baseball was chosen that had a property (in this case innings) which was clearly identifiable with the instance and not the superset. To generate a P1 or P2 sentence, we took a salient property of the superset that could be expressed with the restriction to "is," "has," or "can." In the first example of Table 1, rules were felt to be a very salient property of games. Then an instance was chosen, in this case badminton, to which the P1 property seemed not particularly associated. Our assumption was that, if the model is correct, a typical *S* would decide whether badminton has rules or not by the path, badminton is a game and games have rules.

In Expt 1, false sentences were divided equally between supersets and properties. No systematic basis was used for constructing false sentences beyond an attempt to produce sentences that were not unreasonable or semantically anomalous, and that were always untrue rather than usually untrue. In Expt 2, additional restrictions were placed on the false sentences. The properties of the false PO sentences were chosen so as to contradict a property of the instance itself. In example 3 of Table 1, "Coca-cola is blue" contradicts a property of Coca-cola, that it is brown or caramel-colored. In contrast, the properties of false P1 sentences were chosen so as to contradict a property of the superset. In the same example, alcoholic was chosen, because it is a contradiction of a property of soft drinks in general. The relation of elements in the false SO and S1 sentences can be illustrated by reference to Fig. 1. The false SO sentences were generated by stating that one instance of a category was equivalent to another, such as "A canary is an ostrich." The false S1 sentence was constructed by choosing a category one level up from the instance, but in a different branch of the structure, such as "A canary is a fish."

The sequence of sentences the *S* saw was randomly ordered, except for the restriction to four hierarchies in each run. The runs were counterbalanced over *S*s with respect to the different sentence types, and each button was assigned true for half the *S*s, and false for the other half.

RESULTS AND DISCUSSION

In analyzing the data from the three experiments, we have used the mean RT for

each *S*'s correct responses only. Error rates were on the average about 8% and tended to increase where RT increased.

Deciding a Sentence is True

The data from all three experiments have been averaged in Fig. 2. To evaluate the differences shown there for true sentences, two separate analyses of variance were performed: One for the 2-level runs and one for the 3-level runs. For the 2-level data the difference between P sentences and S sentences was significant, $F(1, 60) = 19.73, p < .01$, the difference between levels was significant $F(1, 60) = 7.74, p < .01$, but the interaction was not quite significant, $F(1, 60) = 2.06$. For the 3-level data, the difference between P and S sentences was significant, $F(1, 60) = 27.02$, $p < .01$, the difference between levels was significant, $F(2, 60) = 5.68$, $p < .01$, and the interaction was not significant, $F < 1$.

Our prediction was that the RT curves for PO, P1, and P2 sentences and for SO, S1, and S2 sentences should be two parallel straight lines. The results are certainly compatible with this prediction, except for the SO point, which is somewhat out of line. It was anticipated that presenting the entire sentence on the CRT at one time would permit the *S*s to answer the SO sentences, e.g., "A maple is a maple," by pattern matching. That they did so was substantiated by spontaneous reports from several *S*s that on the SO sentences they often did not even think what the sentence said. Overall, the underlying model is supported by these data.

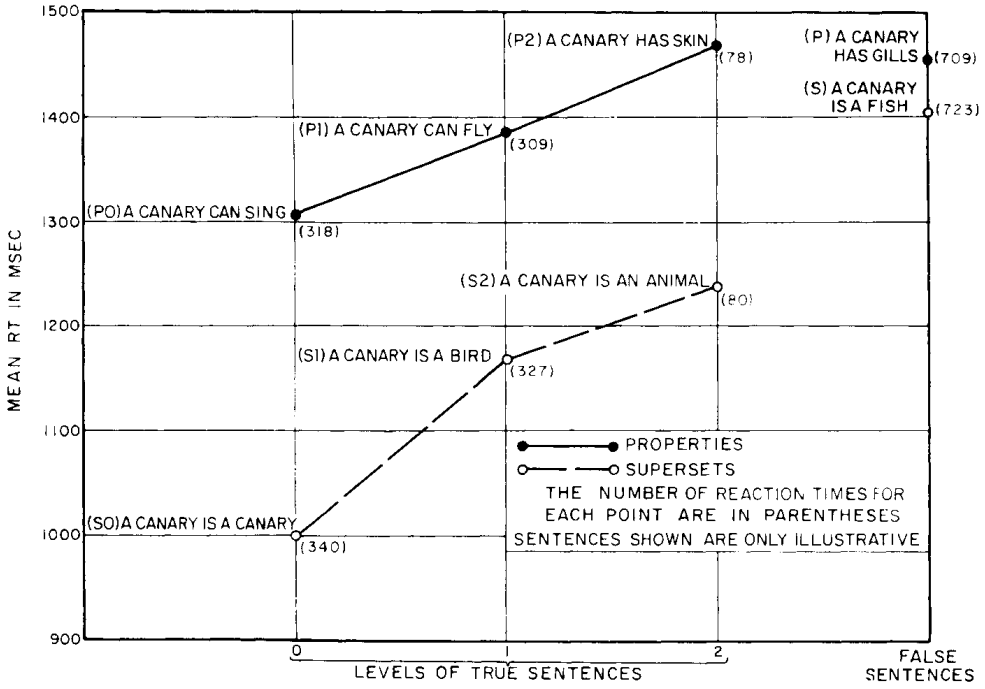

FIG. 2. Average reaction times for different types of sentences in three experiments.

It can also be concluded, if one accepts the model and disregards the SO point as distorted by pattern matching, that the time to move from a node to its superset is on the order of 75 msec, this figure being the average RT increase from PO to P1, P1 to P2, and S1 to S2. The differences between S1 and P1 and between S2 and P2, which average to about 225 msec, represent the time it takes to retrieve a property from the node at the level where we assume it is stored.

We have assumed that retrieval of properties at a node and moving up to the superset of the node are parallel processes, but this was not a necessary assumption. In actual fact the computer realization of the model completes the search for properties at a node *before* moving up one level to its superset. If the property search is assumed to be complete before moving up to the next level, then the 75 msec would have to be divided into two processes: (a) The time spent searching for properties, and (b) the time to move up to the superset. If such an assumption is made, then there is no clear prediction as to whether the increases for P sentences should parallel the increases for S sentences. If, given an S-type sentence, the *S* could dispense with process (a) above, then the slope of the curve for S sentences would be less than for P sentences; if he could not, then the prediction of two parallel lines would still hold. However, the fact that the time attributable to retrieving a property from a node is much longer than the time to move from one node to the next suggests that the processing is in fact parallel. It is unlikely that a search of all the properties at a node could be completed before moving up to the next level in less than 75 msec, *if* it takes some 225 msec actually to retrieve a property when it is found at a node. This might be reasonable if most of the 225 msec was spent in verification or some additional process necessary when the search at a node is successful, but attributing most of the 225 msec to such a process involves the unlikely assumption that this process takes much longer for P sentences than for S sentences. If it were the same for both sentence types, then it would not contribute to the difference (the 225 msec) between their RTs.

Since any other systematic differences between sentence types might affect RTs, we did three further checks. We computed the average number of letters for each sentence type and also weighted averages of the word-frequencies based on the Thorndike-Lorge (1944) general count. Then we asked four *S*s to rate how important each property was for the relevant instance or superset, e.g., how important it is for birds that they can fly. In general, we found no effects that could account for the differences in Fig. 2 on the basis of sentence lengths, frequency counts, or subject ratings of importance. The only exception to this is that the higher frequency of superset words such as bird and animal in the predicates of S1 and S2 sentences may have lowered the averages for S1 and S2 sentences relative to those for P sentences.

Deciding a Sentence is False

There are a number of conceivable strategies or processes by which a person might decide a sentence is false. All of these involve a search of memory; they fall into two classes on the basis of how the search is assumed to terminate.

The Contradiction Hypothesis. Under this hypothesis, false responses involve finding a contradiction between information stored in memory and what the statement says. For example, if the sentence is "Coca-cola is blue," the *S* searches memory until he finds a property of Coca-cola (that it is brown or caramel colored) which contradicts the sentence.

The Contradiction Hypothesis was tested by the construction of false sentences for Expt 2. We predicted that the RT increase from PO to P1 found for true sentences might also be found for false sentences. The difference found was in the right direction, but it was negligibly small (7 msec). Similarly, it was thought that if *S*s search for a contradiction, false SO sentences should produce faster times than the false S1 sentences since there is one less link in the path between the two nodes for an SO sentence. (This can be seen by comparing the path in Fig. 1 between canary and ostrich as in SO sentences to the path between canary and fish as in S1 sentences.) The difference turned out to be in the opposite direction by 59 msec on the average, $t(7) = 2.30$, $p < .1$. If anything, one should conclude from the false SO and S1 sentences in Expt 2 that the closer two nodes are in memory, the longer it takes to decide that they are not related in a stated manner.

The Unsuccessful Search Hypothesis. This is a generalization of what Sternberg (1966) calls the "self-terminating search," one of the two models he considered with regard to his RT studies of short-term memory search. Under this hypothesis an *S* would search for information to decide that a given sentence is true, and, when the search fails, as determined by some criterion, he would respond false. One possible variation, suggested by the longer RTs for false responses, would be that *S*s search memory for a fixed period of time, responding true at any time information is

found that confirms the statement is true, and responding false if nothing is found by the end of the time period. Such a hypothesis should lead to smaller standard deviations for false sentences than for true sentences, but the opposite was found for Expt 2, where it could be checked most easily.

The Search and Destroy Hypothesis. We developed another variation of the Unsuccessful Search Hypothesis after the Contradiction Hypothesis proved unsatisfactory and *S*s had been interrogated as to what they thought they were doing on false sentences. Under this hypothesis we assume the *S* tries to find paths through his memory which connect the subject and predicate of the sentence (e.g., the path "canary → bird → animal → has skin" connects the two parts of "A canary has skin"). Whenever he finds such a path he must check to see if it agrees with what is stated in the sentence. When the *S* has checked to a certain number of levels or "depth" (Quillian, 1967), all connections found having been rejected, the *S* will then respond false. Under this hypothesis, the times for false sentences will be longer, in general, and highly variable depending upon how many connective paths the *S* has to check out before rejecting the statement. For instance, assuming people know Coca-cola comes in green bottles, a statement such as "Coca-cola is blue" would on the average take less time than "Coca-cola is green." This is because the *S* would have to spend time checking whether or not the above path between Coca-cola and green (i.e., that its bottles are green) corresponds to the relation stated in the sentence.

This hypothesis would explain the longer times in Expt 2 for sentences such as "A canary is an ostrich" as compared with "A canary is a fish" in terms of the greater number of connections between canary and ostrich that presumably would have to be checked out. This difference in the number of connections would derive from the greater number of properties that are common to two nodes close together in the network, such as canary and ostrich, than are common to nodes further apart and at different levels, such as canary and fish.

Finding contradictions can be included in this hypothesis, as is illustrated with "Gin is wet." Here the *S* might make a connection between gin and wet through the path "gin is dry and dry is the opposite of wet." Seeing the contradiction, he rejects this as a basis for responding true, but continues to search for an acceptable path. In this example, if he searches deep enough, he will find the path "gin is liquor, and liquor is liquid, and liquid is wet" which is, in fact, what the sentence requires. The point we want to emphasize here is that even though a contradiction can be used to reject a path, it cannot be used to reject the truth of a statement.

There are certainly other possible hypotheses, and it is possible that a combination of this hypothesis with the Contradiction Hypothesis may be necessary to explain false judgments. Needless to say, the process by which a person decides that a statement is false does not seem to be very simple.

Conclusion

In a computer system designed for the storage of semantic information, it is more economical to store generalized information with superset nodes, rather than with all the individual nodes to which such a generalization might apply. But such a storage system incurs the cost of additional processing time in retrieving information. When the implications of such a model were tested for human *S*s using well-ordered hierarchies that are part of the common culture, there was substantial agreement between the predictions and the data.

There is no clear picture that emerges as to how people decide a statement is false. Our current hypothesis, that people must spend time checking out any interpretations that are possible (see the discussion of the Search and Destroy Hypothesis), should be testable,

but even corroborative evidence would not clear up many of the questions about such decisions.

The model also makes predictions for other RT tasks utilizing such hierarchies. For instance, if *S*s are given the task of deciding what common category two instances belong to, then RT should reflect the number of supersets the *S* must move through to make the decision. (Consider fish and bird, vs. shark and bird, vs. shark and canary; see Fig. 1). Such RT differences should parallel those in our data. Furthermore, if utilizing a particular path in retrieval increases its accessibility temporarily, then we would expect prior exposure to "A canary is a bird" to have more effect in reducing RT to "A canary can fly" than to "A canary can sing." There are many similar experiments which would serve to pin down more precisely the structure and processing of human semantic memory.

References

QUILLIAN, M. R. Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Sci.*, 1967, **12**, 410–430.

QUILLIAN, M. R. The Teachable Language Comprehender: A simulation program and theory of language. *Communications Assn. Comp. Mach.*, 1969, (In press).

SMITH, E. E. Effects of familiarity on stimulus recognition and categorization. *J. exp. Psychol.*, 1967, **74**, 324–332.

SMITH, E. E. Choice reaction time: An analysis of the major theoretical positions. *Psychol. Bull.*, 1968, **69**, 77–110.

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

THORNDIKE, E. L. AND LORGE, I. *The teacher's word book of 30,000 words.* New York: Columbia Univ. Press, 1944.

(Received July 31, 1968)