COGNITIVE PSYCHOLOGY 1, 59–85 (1970)

Processes in Word Recognition[1]

DANIEL D. WHEELER[2]

Mental Health Research Institute, The University of Michigan

Five hypotheses were proposed and tested to account for Reicher's (1968) finding that recognition of letters is more accurate in the context of a meaningful word than alone, even with redundancy controlled by a forced-choice design. All five hypotheses were rejected on the basis of the experimental results. Performance on the forced-choice letter detection task averaged 10% better when the stimuli were four-letter English words than when the stimuli were single letters appearing alone in the visual field.

Three classes of models were proposed to account for the experimental results. All three are based on analysis of the task in terms of the extraction of features from the stimuli.

This decade has provided a number of advances in our understanding of how humans process visual information. On the experimental side, Sperling (1960) and Averbach and Coriell (1961) have demonstrated the existence and general nature of the icon (Neisser, 1967). Recoding from the icon was shown by Conrad's (1964) acoustic confusability results. And the parallel versus serial issue in the processing of visual arrays has been investigated by Estes and Taylor (1964, 1966) and Wolford, Wessel, and Estes (1964). The theoretical advances have included the development of a number of pattern recognition models (Uhr, 1966; Selfridge, 1966). Some, like Rumelhart (in press), are beginning to fit models to detailed experimental data. The stimuli used in most of this work were letters, either singly or in arrays, rather than words.

[1] This article is a modified version of a Ph.D. dissertation submitted to The University of Michigan. I would like to thank my Committee, Robert A. Bjork, James G. Greeno, and Wilfred M. Kincaid, and especially my Chairman, Walter Reitman, for their help and encouragement. I am indebted to Jonathan Baron, Donald Broadbent, Jerry Gardner, George Mandler, Arthur Melton, and Donald Norman for comments and discussion. The late Peter Headly deserves considerable credit for programming the experimental system. Mrs. Billie Lawson and Miss Judith Brunclik helped devise the stimulus materials. Mrs. Joan McClain did a wonderful job of typing everything through at least three versions of the paper. This work was supported by USPHS Grant No. MH12160, which I should like to acknowledge with appreciation.

[2] Now at the University of Texas at Austin.

The major line of recent work on word recognition has been fairly independent of the other work on visual information processing. It traces back to Howes and Solomon's (1951) demonstration that the visual duration thresholds for words are a function of word frequency. There has been a major theoretical controversy over whether the word frequency effect reflects some basic property of the perceptual mechanism or whether the effect is attributable to a response bias from the subjects' greater tendency to use high frequency words. Broadbent (1967) and Morton (1968) have tested a number of specific models from two classes, guessing models and signal detection models. In general, their analyses of data from auditory perception of words in noise support a signal detection model with a criterion shift towards more frequent words.

A tacit assumption common to both the work with letter arrays and the models for the word frequency effect is that the perceptual aspects of word recognition can be understood in terms of individual letter recognition. Only very general interactions among letters should occur from changes in attention or overall contrast. Other than these effects, perception of one letter (or extraction of information from the letter) should be independent of perception of the others. The effects of set, word frequency, etc., are introduced by a decision mechanism which takes advantage of the redundancy of words.

This independence assumption is challenged by Reicher's (1968) study of word recognition. Reicher probed the accuracy of the recognition of the individual letters of a word in a tachistoscopic exposure by giving the subject a forced-choice test between two letters, one of which appeared in the stimulus. Normally the redundancy of English words would prevent this probe technique from being specific to a single letter of the stimulus because the identity of a letter can often be inferred from the other letters of the word. Reicher eliminated the effects of redundancy by having both alternatives form a word with the remaining letters. For example, *D* and *K* might be the alternatives for testing the fourth position of the stimulus *WORD*. The untested three letters *WOR* should contribute no information to the choice between *D* and *K* since *WORK* is also a common English word.

Reicher's main finding was that performance on the forced-choice tests of letter recognition was more accurate when the stimulus was a four-letter word than when it was either a single letter or a nonsense quadrigram. Since the subjects did not know until after the presentation of the stimulus what letter position would be tested, the strong word superiority effect means that subjects recognized all four letters of the word with a higher probability of being correct than they had on a single letter alone.

Figure 1 shows examples of the experimental materials used by

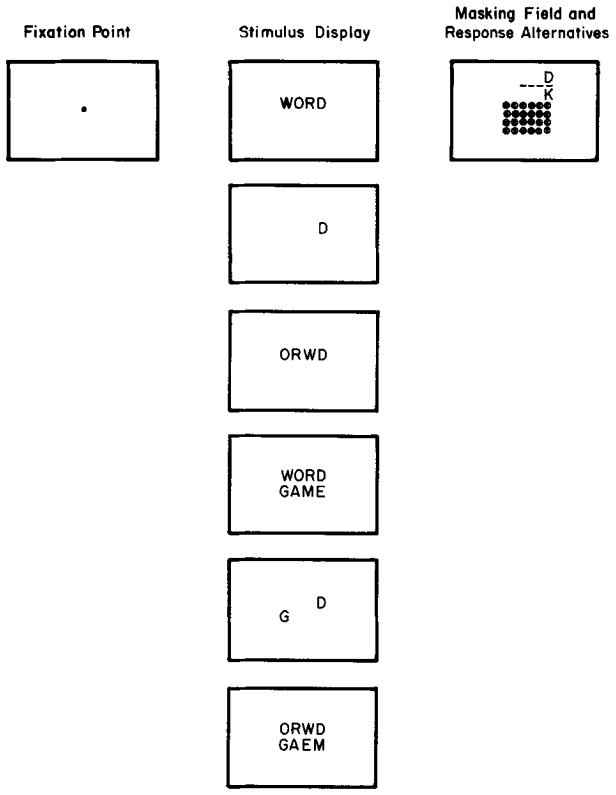

FIG. 1. Examples of tachistoscopic displays used by Reicher (1968, Figure 2). The stimulus display always consisted of either one or two stimuli of the same type, words, letters, or quadrigrams.

Reicher. On each trial the fixation point was displayed until the subject initiated the stimulus exposure. One of the six types of stimulus displays was then briefly exposed and then immediately replaced by the masking field with the alternatives. The position of the alternatives marked the position of the probed letter in the word.

Reicher found that performance, as measured by the proportion of correct choices, was better with word stimuli than either single letters or quadrigrams by about 8%. The results held with both single and double stimuli over a range of exposure times. Performance on the quadrigrams was worst, but apparently not significantly worse than on single letters.

Reicher also found that performance was consistently worse in a pre-cue condition, where subjects were told verbally before each trial what the alternatives would be. The difference was about 8%. There was no other change in the procedure; the alternatives also appeared visually

along with the mask exactly as they did in the no pre-cue condition. The results are directly contrary to the theory that set or expectation for the stimulus to be presented should improve performance.

Implications of Reicher's Results

At first glance, Reicher's experimental paradigm appears to provide a test of the serial versus parallel processing issue in the organization of visual information processing in humans. The simple versions of parallel and serial models make different predictions about the results of Reicher's experiment. Most serial processing models for the readout of information from the visual image or icon would predict that subjects in Reicher's experiment should do best on the single letter stimuli. With more than one letter to process, the average amount of processing per letter during the limited iconic duration must decrease.

The simple parallel models usually assume that the letter units are processed independently at the same time. There should be no difference in the processing of single letters alone and in a word. Performance in Reicher's experiment should be the same with letter and word stimuli.

Neither prediction is consistent with the obtained results. Neither model has any mechanism that would explain the superiority of performance on words. As soon as a mechanism is added to account for more accurate performance on meaningful words, either model can fit the observed data.

There are two ways around this theoretical impasse. One can hypothesize separate factors or processes which account for the superiority of words without changing the basic pattern recognition models. These are generally consistent with either parallel or serial models and can be included in the models without drastic changes. Or one can modify the basic pattern recognition models by dropping the independence assumption and proposing some interactive recognition system in its place. The first approach leads to testable hypotheses and simpler models. It will be attempted first.

Hypotheses for Separate Mechanisms

The five hypotheses discussed below all suggest mechanisms or processes that could be included in either a serial or parallel model to account for the word superiority effect.

1. *Interference hypothesis.* In Reicher's experiment, the presentation of the stimulus display was terminated by the onset of a field containing both the mask and the two choice alternatives. The subject than had to recognize the two choice letters before he could decide between them. Some of Reicher's subjects reported that the choice letters interfered with

the recognition of the stimulus. Reicher argued that the recognition of the single letter choice alternatives would interfere with the recognition of the single letter stimuli more than with the recognition of the word stimuli. This would be true if, for instance, the interference occurred in that part of the pattern recognition process which involved access to memory to find the stored representation of the object identified. Such interference would result in better performance on words.

This hypothesis can easily be tested. The interference should not occur if the recognition of the alternatives is separated in time from the recognition of the stimulus. This can be accomplished by delaying the presentation of the alternatives until the recognition of the stimulus is completed. Thus if the hypothesis is correct the difference between the performance on words and letters should disappear when the presentation of the choices is delayed sufficiently.

2. *Preprocessing hypothesis.* A number of pattern recognition models have postulated a stage of preprocessing which comes before the actual pattern recognition process (see Uhr, 1966). This stage is supposed to isolate and normalize the size, position, etc., of the stimulus to be recognized. In order to control for differential sensitivity within the foveal area, Reicher presented his single letter stimuli at the same position they would have appeared at had they been in the corresponding word. Since the words were centered with respect to the fixation point, the positions of the single letter stimuli had to vary. This might cause the preprocessing stage to take longer to isolate the single letter stimuli in the visual field. Thus there would be less time remaining before the icon faded for the actual pattern recognition system to work on the single letter stimuli.

The positional uncertainty of the single letter stimuli can be eliminated by positioning the single letter stimuli at a constant position in the visual field defined by the fixation point. If the superiority of performance on word stimuli is a result of the additional preprocessing required to locate letter stimuli in the visual field, the effect should disappear when the positional uncertainty of the letters is eliminated. This manipulation does not, however, control for variation in foveal sensitivity. As a further check on a possible preprocessing stage and the effects of positional uncertainty, another condition could be run so the words rather than the letters have positional uncertainty. Each word could be presented with the letter being tested positioned at the fixation point. Thus the position of the whole word would vary as a function of the position of the letter being tested. When compared with single letters in the constant position, performance on words should be worse than on single letters if the preprocessing and positional uncertainty hypothesis is correct.

3. *Focusing hypothesis.* There may be some idiosyncratic properties

of individual words that cause the pattern recognizing mechanism to focus on those aspects of a word which contain the most information that distinguishes the presented word from other words. These aspects are likely to be in those letter positions for which there are alternative letters that can be switched to form similar words. This is, of course, exactly where the stimuli were tested in Reicher's experiment. The appropriate control is to choose the stimulus words so that they have an alternative for every letter position. For example, *READ* can be changed one letter at a time to form *HEAD, ROAD, REND,* and *REAL.* If the effect holds over all positions in the same word, it is hard to see how a focusing mechanism could account for Reicher's results.

4. *Response bias hypothesis.* If subjects see some of the letters of a word stimulus, but not the letter tested, they are probably more likely to guess the alternative that forms the more frequent word with those letters he has recognized. Reicher does not mention any control of word frequency. The simplest control for a possible response bias effect of word frequency is to use both words as stimuli. If one subject gets the stimulus *READ* with choices *R* and *H*, another subject should get *HEAD* with the same choices. Any improvement in performance on the more frequent words will be cancelled by poorer performance on the other words.

5. *Word frequency hypothesis.* The effects of word frequency need not be limited to a guessing process that would produce a response bias. Recognition involves the access to the subject's long-term memory for the object being recognized. Access may be easier for more frequent words. Thus performance on the TAME/TAMP pair would be worse than on the pair of more frequent words CARE/CAKE.

Single letters can be considered as low frequency words. The letter *E* is the most frequent letter in English, but appears alone as a unit only in such unusual contexts as "row *E*" on a theater ticket. Thus a word frequency effect might explain the superiority of performance on words.

The crucial data for a test of this hypothesis is the performance on the single letters *I* and *A*. These letters are also high frequency English words. Both are among the 500 most frequent words in the Thorndike-Lorge (1944) count. If the superiority of performance on words is attributable only to the fact that they appear more frequently on the average than the letters appear as units, then the performance on the single letters (words) *I* and *A* should be as good as the performance on high frequency words.

EXPERIMENTAL METHOD

The tests of the five hypotheses were carried out in a single experiment. The experimental paradigm was very similar to that used by Reicher (1968) in his no pre-cue conditions. Two features of Reicher's stimulus

displays, however, seem irrelevant in the light of his results. The first is the use of either one or two rows of stimuli. Reicher's main results held for both single and double stimuli. The second feature is the use of the position of the alternatives to mark the position of the critical letter of the stimulus. This makes Reicher's experiment similar in some respects to Sperling's (1960) partial report paradigm. But if Reicher's mask field is effective, the position cue should be of little use to the subject. Both of these apsects of Reicher's experiment were eliminated in the experiment designed to test the five hypotheses.

Apparatus and Experimental Setting

The experimental equipment consisted of a Digital Equipment Corporation PDP-8 computer connected to a Tektronix Model 611 storage scope. The computer was programmed to display a series of four visual fields on the scope, making it much like a four-channel tachistoscope.

The display scope and response keys were in an experimental room separate from the computer and teletype. The subjects were run individually, seated in front of a panel on which the response keys were mounted. The display scope was positioned at eye level on a wheeled cart behind the panel. This allowed the distance from the subject to be adjusted so that the mask field subtended an angle of $2\frac{1}{2}$ degrees for each subject. A $7\frac{1}{2}$ watt nightlight provided dim overall illumination for the experimental room, sufficient for the dark-adapted subjects to see the response keys, yet not so much that the subjects could see their reflection in the glass scope face. An intercom allowed the experimenter in the computer room to communicate with the subject.

The stimulus material was read into the computer from paper tape during the course of the experiment. For each trial the following were read in: (*a*) a seven-letter stimulus field (three of which would be blanks for four-letter words), (*b*) the correct and incorrect choice letters, (*c*) a stimulus display time code, (*d*) a choice delay time code, and (*e*) a category code for data analysis. At the beginning of each experimental run, the program requested stimulus display times and choice delay times for each of the codes used on the tape of items. The display times were in units of 5 msec and the choice delay times were in seconds.

Figure 2 shows the physical arrangement of the display fields. The sequence of events during each trial was as follows: First the fixation point was displayed. Then the subject initiated the trial by pressing a key with his left hand. The fixation point disappeared, and the stimulus field was displayed for the length of time specified by the code on the item tape and the correspondence set up for that code at the beginning of the run. Immediately after the offset of the stimulus the mask field was displayed.

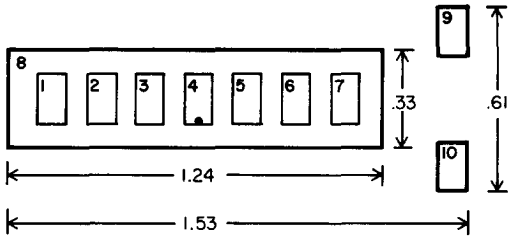

FIG. 2. Arrangement of the display fields on the face of the scope. The dimensions are in inches. The letter position areas are .11 × .16 in. Field 1: The fixation point appeared at the bottom of area 4. Field 2: The stimulus appeared in letter positions 1–7. Field 3: A masking pattern of random dots covered area 8. Field 4: The choices were displayed in letter positions 9 and 10 as the masking pattern remained on.

The mask field consists of a random pattern of dots, different on each trial. After the choice delay interval (possibly 0 sec), the two choices appeared to the right of the mask field. Both the mask and the choice fields remained on until the subject pressed one of the response keys. The average rate was approximately one trial per $5\frac{1}{2}$ sec.

One disadvantage of the computer-operated experimental apparatus was that the brightness of the display scope could not be controlled precisely. The brightness adjustment on the scope provided only coarse control; small changes in the knob position caused large changes in the brightness. Furthermore, the brightness of the scope drifted over time. Fortunately, the drift was slow and of relatively small magnitude.

Materials

The test of the focusing hypothesis requires that we use words that can be tested in every letter position, i.e., words that have an alternative that forms another meaningful word in each letter position. These words will be called base words. The test of the response bias hypothesis requires that the alternative for every test also be tested. Thus the stimulus words were made up in sets of five, one base word and four alternatives. An example of a base word is *READ*, with the alternatives *HEAD*, *ROAD*, *REND*, and *REAL*.

Forty-eight nonoverlapping sets of five words were found with the aid of a crossword puzzle dictionary of four-letter words. Each set of five words provided 16 test items: four tests of the base word (once in each letter position); four tests of the alternative words; and eight single letter items constructed from the word items. The single letter items were formed by removing the three untested letters of the word items. The single letter remaining was used with the same alternatives as the original word. Table 1 shows how the 16 items were formed and divided into

groups. It would not be wise to give the same subject all 16 items, especially the four tests of the base word. Thus the items were divided into four groups. Into each group went a test of one of the positions of the base word, the test of the alternative formed from a different position of the key word, and two single letter items. Each group was balanced so that each letter position was tested equally often on both base words and alternatives. Since there were 48 different base words, a complete group consisted of 192 (4×48) items.

Another group of 192 items was constructed for use in estimating the exposure duration required to obtain the desired percentage correct for each subject. This group also consisted of half words and half single letters. All four-letter positions were tested equally often, and none of the words used overlapped those in the fully balanced groups. A small group of 20 items was constructed to serve as examples during the instructions.

The test of the preprocessing hypothesis requires different ways of placing the words and letters into the seven positions of the display field (see Figure 2). Both words and letters could vary in position or remain constant. When word position was constant, the four letters were placed

TABLE 1
Construction of 16 Items from a Single Base Word, *READ*

Type	Stimulus	Choices		Stimulus group
		Correct	Incorrect	
4 Tests of base word	READ	R	H	1
	READ	E	O	2
	READ	A	N	3
	READ	D	L	4
Alternative words	HEAD	H	R	3
	ROAD	O	E	4
	REND	N	A	1
	REAL	L	D	2
Single letter items from base word	R	R	H	2
	E	E	O	3
	A	A	N	4
	D	D	L	1
Single letter items from alternative words	H	H	R	4
	O	O	E	1
	N	N	A	2
	L	L	D	3

Note: Items are divided into four groups so that each subject is tested on only one item of each type.

TABLE 2
Arrangement of the Stimuli in the Display Field for the Three Conditions

Letter position tested	Condition: Letter shift	Condition: No shift	Condition: Word shift	Correct choice	Incorrect choice
1	HEAD H	HEAD H	HEAD H	H	R
2	ROAD O	ROAD O	ROAD O	O	E
3	REND N	REND N	REND N	N	A
4	REAL L	REAL L	REAL L	L	D

Note: When words were shifted, the letter tested was always placed at the fixation point. When letters were shifted, they were placed in the same position in which they had appeared in the word from which they were derived.

in positions three through six. When the word position was varied, it was placed in the field so that the tested letter appeared in position four. Letters in constant position appeared in position four. When letters varied, they appeared in the position in which they would have been in the word from which they were derived (see Table 2).

Separate item tapes were made for each group of stimuli in the following conditions: (*a*) letter shift, with letter position varying and word position constant, (*b*) word shift, with word position varying and letters constant, and (*c*) no shift, with neither position varying. Within each of the first two conditions, the distribution of foveal positions of the letters tested in a word and the single letters are the same. The effect of position uncertainty is contrasted across the two conditions. The foveal position is not balanced in the no-shift condition, but the position uncertainty is the same for letters and words.

The letter or word shift variable was the only difference among the three experimental conditions. The other independent variables were manipulated within subjects. Three values of choice delay time were used to provide the test of the interference hypothesis. The time between the stimulus offset/mask onset and the presentation of the choice letters was 0, 1, or 2 sec.

The independent variables within each condition were completely counterbalanced so that each level of each variable was tested equally often with each possible combination of levels of the other variables. The balanced variables were the following: (*a*) word or letter stimulus, (*b*) choice delay time, (*c*) base word or alternative word, and (*d*) letter position of the tested letter. There are $2 \times 3 \times 2 \times 4 = 48$ combinations; each subject was tested four times on each combination for a total of 192 trials.

The 192 items were divided into two matched sets of 96. The items were randomized within each set of 96. Twelve different randomized orders were used, four in each condition.

Procedure

Each subject was run for one experimental session lasting approximately 1 hr. The session began with an instructional group of 20 trials, with display times beginning at $\frac{1}{2}$ sec and decreasing to 25 msec. As the subject worked through these items, the experimenter explained the procedure. The following points were emphasized: (*a*) the stimulus would always be a single letter or a four-letter English word, (*b*) they were to perform as accurately as possible and guess when necessary, and (*c*) they were to work at a rate they found comfortable. The experimenter left the room as the subject finished the instructional trials.

The subject then worked through four sets of 96 items, with a 2-min break between each set. The exposure duration was varied in the first two sets. Performance on these sets was used to estimate the exposure duration which would result in 75% correct performance on the remaining sets. The last two sets of 96 were from the matched groups of base words and alternatives. Only the data from the last two sets were used in the complete analysis. The exposure time for the fourth set was adjusted if performance on the third set was not in the range of 15 to 39% errors. An adjustment was necessary for 13 of the 36 subjects.

At the conclusion of the experiment, the subjects were interviewed to obtain their general reactions, to check on the strategies the subjects may have used, and to test their awareness of the positional shifts (if any) in their condition.

Subjects

Subjects were 36 paid volunteers from the Mental Health Research Institute subject pool. Most were college students. Each served for one experimental session, approximately 1 hr long. Six male and six female subjects served in each of the three conditions. Subjects were assigned to conditions by a rotating scheme in the order in which they participated in the experiment.

RESULTS

Overall Results

Performance on words was consistently better than performance on single letter items in all three conditions. The average difference in performance, in favor of words, is 10%. Table 3 shows both the overall results and the breakdown by condition. Considering individual subjects, only four failed to perform better on words than on letters. One had equal performance on letters and words, two missed only one more word than letter, and one made three more errors on words. All four of these subjects were female.

Three of these four subjects were in the same condition, the letter shift condition. But the overall superiority of performance on words within that condition was still large. A check using Kincaid's (1962) method of combining contingency tables[3] showed that the difference between the performance on words and letters within the letter shift condition was significant at the .001 level. The larger differences in the other groups were significant beyond the .001 level.

The apparent differences among the conditions in both overall level of performance and the size of the difference between performance on letters and words are probably artifactual. The exposure time for each subject was determined from performance on the first two sets of 96 items by estimating the exposure time that would produce performance closest to an overall 75% correct level. The observed differences among the conditions reflect failures of the estimation procedure, complicated by the lack of fine control over brightness and exposure duration, to obtain the 75%

TABLE 3
Percent Correct Responses by Condition

Condition	Percent correct		Word-letter difference
	Words	Letters	
Letter shift	72.5	65.8	6.7
No shift	80.8	67.5	13.3
Word shift	74.9	65.0	9.9
Mean	76.1	66.1	10.0

Note: 1,152 observations per entry.

[3] All significance levels reported in this work were obtained by this method unless otherwise stated. A sign test (Siegel, 1956) was used only when Kincaid's method was difficult to apply.

performance level. The mean exposure times for the three groups vary in exactly the order as the overall level of performance; the subjects in the no shift condition has both the largest average exposure time and the highest percentage correct on both letters and words.

The differences in the relative advantage of words over letters among the three conditions remain to be explained. Were the exposure level set so that performance approached either chance or perfect, the word-letter performance difference obviously would disappear. At some intermediate point, halfway between perfect and chance performance by most simple choice theories, the effect should be maximum. The order of the size of the word-letter effect is indeed exactly the same as the order of the overall performance within the conditions away from the 75% level. The subjects in the no shift condition, with overall performance closest to 75%, showed the largest difference between the performance on letters and words.

The above arguments suggest that the apparent differences among the conditions in both performance level and magnitude of the word superiority effect are probably artifactual. The only firm conclusion possible is that there is a sizeable word superiority effect in all three conditions.

Tests of Specific Hypotheses

1. *Interference hypothesis.* Within each condition, three delay times (0, 1, and 2 sec) were used between the onset of the mask and the presentation of the choices. If the process of recognizing the alternatives interferes with the still proceeding process of recognizing the stimulus when the delay time is zero, the interference should be reduced and performance should improve when the delay times are longer. Furthermore, the improvement should be greater for letters than for words because the single letter choices are more similar to the single letter stimuli than to the word stimuli.

The data shown in Figure 3 confirm this prediction. For the letter items, the delay time had a significant (.001 level) effect on performance. The percent correct was lowest at the zero delay interval, where the interference should have been strongest. Performance on the word items increased slightly as the delay time increased, but the increase was not significant ($.10 > p > .05$).

Despite this confirmation of the interference effect, that effect is not sufficient to account for the word superiority effect. As Figure 3 clearly shows, there was still a considerable difference in performance in favor of the words beyond the zero delay point. Within the 1 and 2 sec delay items, the mean percentages of correct responses were 77.1% for words and 69.1% for letters. The difference of 8.0% is significant at the .001 level.

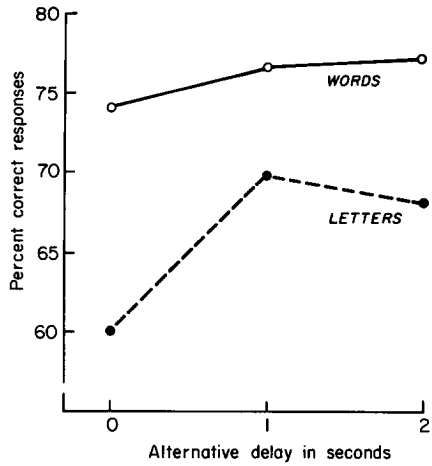

FIG. 3. Percent correct by alternative delay for letter and word items.

The word superiority effect with choice delays of 1 and 2 sec was significant within each of the three conditions. This eliminates the possibility that the interference effect interacting with the conditions could explain the word superiority effect. The smallest difference among the conditions was 5.2%, significant at the .025 level.

The comments of subjects in the postexperimental interview suggest a more specific explanation of the nature of the interference effect. Several subjects reported that frequently the letter stimuli seemed to jump from the stimulus position to one of the alternative positions when there was zero delay. This is, of course, the standard apparent motion phenomenon. If the apparent motion were consistently to the correct alternative, it would improve performance on the single letters in the zero delay condition. On the other hand, if the direction of the apparent motion were not related to the position of the correct alternative, the apparent motion might interfere with the identification of the stimulus. The latter possibility would explain the obtained effect.

2. *Preprocessing hypothesis.* If the variance in the letter position causes the word superiority effect by slowing down a preprocessing mechanism in the human pattern recognition system, the effect should disappear when the variance in the letter position is eliminated. Thus the hypothesis predicts that the word superiority effect should appear in only the letter shift condition of the present experiment. In fact, performance on letters was poorer than on words in the other conditions as well, even though the letters appeared in a constant position in those conditions. The preprocessing hypothesis can be rejected.

3. *Focusing hypothesis.* The operation of a focusing or attention directing mechanism should not improve the forced-choice performance on word items when each word is tested in all four letter positions. Thus if the hypothesis is correct, the word superiority effect should not be found with base words (words tested in all positions) in the present experiment. The word superiority effect should be found only with the alternatives to the base words.

Table 4 shows the percentage correct for the base words and alternative words. The difference between them is not significant. Performance on both base words and alternative words is significantly (.001 level) better than performance on single letters, even by a simple sign test. The focusing hypothesis can be rejected.

4. *Response bias hypothesis.* If the word superiority effect is simply the result of a response bias established by the untested letters of the words towards the correct alternative, the effect should be reversed when the stimuli are changed so that the other alternative becomes correct. The net effect in favor of words should disappear when both alternatives are tested, as in this experiment. The strong word superiority effect shown in the overall results (see Table 3) demonstrates that the response bias hypothesis cannot be the correct explanation.

5. *Word frequency hypothesis.* The critical data for a test of the hypothesis that the word superiority effect is due to the higher frequency of words as units than letters as units is the performance on the single letters *I* and *A*. Both of these are also high frequency words, within the top category in the Thorndike-Lorge (1944) count. Table 5 shows the percentage correct for *I*'s and A's as single letters and for words in which the letters *I* and *A* were tested. In both cases performance on the four-letter words is better than performance on the single letters *I* and *A*. A sign test showed these differences to be significant, at the .002 level for

TABLE 4
Percent Correct Responses for Base Words and Alternative Words

Condition	Percent correct	
	Base words	Alternative words
Letter shift	72.4	72.6
No shift	80.4	81.2
Word shift	73.4	76.4
Overall	75.4	76.7

Note: 576 observations per entry.

TABLE 5
Percent Correct Responses on Tests of *A* and *I*

Letter tested	Context		*N*
	Word	Single letter	
A	78.4	70.1	351
I	80.9	59.7	225
Combined	79.3	65.6	576

A's and beyond the .001 level for *I*'s and for the combined data. The combined performance on *A*'s and *I*'s is slightly worse than the average percent correct for all letters (see Table 3). These results show that the word superiority effect cannot be explained in terms of a word frequency effect.

Serial Position Curves

The serial position curves plotted in Figure 4 show performance as a function of the position within the word of the letter tested. Performance on single letters in the letter shift condition is shown as a function of the position of the letter in the stimulus field. Some aspects of the processing mechanism should influence the shape of these curves. Unfortunately, the effects of irrelevant variables prevent the making of strong inferences about the processes from the shapes of the serial position curves. The decrease in sensitivity away from the center of the fovea predicts decreasing performance as distance from the fixation point increases. Performance at position four was indeed the worst except in the single letter case. Unknown effects are added by the different distributions of letters at each letter position and by the possible interaction of the overall level of performance with the serial position effect.

The difference between the serial position curves for the no shift and letter shift (words) conditions is interesting. The stimuli in these two conditions were exactly the same. The decreasing performance with serial position in the no shift condition is what would be expected from a serial processing mechanism. The flatter curve in the letter shift condition is more consistent with a parallel processing mechanism. It is unlikely that there are separate mechanisms brought into use in each condition. A more natural explanation is that the distribution of the position of the letters in the letter shift condition causes the subjects to spread their attention more evenly over the stimulus field. In the no shift condition $\frac{5}{8}$ of the tests (all single letters and $\frac{1}{4}$ of the words) are in the second letter

position. A concentration of attention at the beginning of the word would result in the observed decrease in performance towards the end of the word.

The serial position curve for the word shift condition is not directly comparable to the others. The position of the tested letter in the stimulus field was always the same, but the other letters of the word varied in position (see Table 2).

The sharp dip in the serial position curve for letters is surprising. No comparable dip was apparent in any of the curves for words. Two possible explanations are interference from the fixation point and lower discriminability of the letters tested at that position (79% vowels). But both of these should affect performance on words as well as the letters. Reicher's (1968) serial position curves for single letters do not show a drop at serial position two when the single letters were in the upper row, but do show a drop in the lower row.

Latency Data

The emphasis in both the theory and design of this experiment was on response accuracy. The instructions to the subjects stressed accuracy,

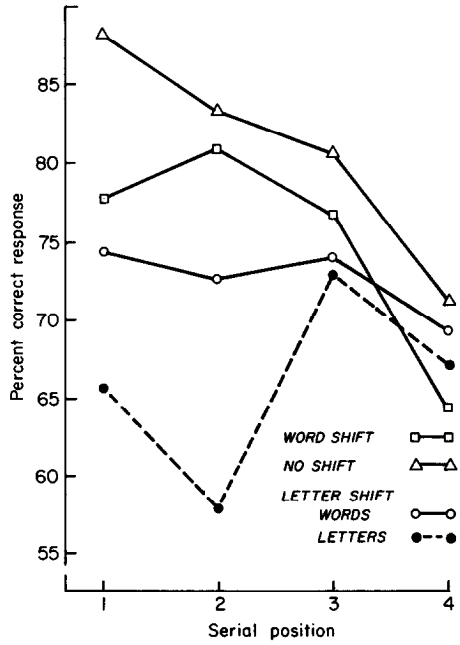

FIG. 4. Serial position curves.

TABLE 6
Mean Latencies of Forced-Choice Responses

	Correct	Incorrect
Words	1228	1679
Letters	1119	1395

Note: Mean latencies in milliseconds averaged over all conditions and alternative delays.

and the fact that response time was being recorded was not mentioned. These facts call for considerable caution in interpreting the latency data.

Table 6 shows the mean response latencies separately for correct and incorrect words and letters. The latencies were longer for words than for single letters, and longer for errors than correct responses. These relationships are reliable; they held within every choice delay in all conditions, a total of nine comparisons.

At first glance these data seem to support the finding of Stewart, James, and Gough (1969) that recognition latency increased as a function of word length. Based on their data, one would expect the recognition time for single letters to be less than that for four-letter words. The forced-choice task, however, may add another process to those required to make a response to a word stimulus. After the subject has seen the word stimulus and the two alternatives, he may have to scan the word in some way to find the letter that is one of the alternatives. This scan is not necessary with single letter stimuli; they can be compared immediately against the two alternatives. Thus the total response times for words and letters, when measured from the onset of the stimulus, are:

$$t_{\text{words}} = t_{\text{wr}} + t_{\text{scan}} + t_{\text{rest}} \quad (1)$$

$$t_{\text{letters}} = t_{\text{lr}} + t_{\text{rest}} \quad (2)$$

where t_{wr} is word recognition time, t_{lr} is letter recognition time, t_{scan} is time to scan through the word, and t_{rest} is the time for all other processes involved, including recognizing the alternatives, making the decision, and making the motor response. It is obvious that the relationship between word and letter recognition times cannot be inferred from the relationship between t_{words} and t_{letters}.

Stewart, James, and Gough (1969) obtained estimates of recognition time alone by subtracting the production latencies from the total response latencies. The production latencies were obtained by letting the subjects recognize the stimuli and then giving them a separate signal to produce the response. Essentially the same thing can be done using the latency data from the 0 and 1 sec alternative delay conditions. When the alter-

natives are delayed for a second, the recognition process should be completed by the time the alternatives are presented, but the scan and all the processes included in t_{rest} cannot begin until alternatives are available. Thus the overall response times with delayed presentation of the alternatives are:

$$t_{words} = t_{scan} + t_{rest} \quad (3)$$

$$t_{letters} = t_{rest} \quad (4)$$

Subtracting the response latencies in the 1 sec delay condition from those with zero delay (i.e., equations (3) and (4) from equations (1) and (2), respectively) produces estimates of word and letter recognition latencies separate from the times for other processes.

Table 7 shows these calculations carried out. The response latencies entered in Table 7 have been corrected for bias from the longer latencies produced by the guessing process using the method of Wolford, Wessel, and Estes (1968). The estimated word recognition time is considerably longer than the letter recognition time. These results are consistent with the finding of Stewart, James, and Gough (1969) that recognition time increases with the length of the word. The same relationship is found when the latencies are not corrected for bias.

The above argument depends critically on a number of untested assumptions about the temporal independence and identification of the processes involved. It is possible that subjects use a direct physical comparison process between the single letter stimuli and the response alternatives. The reports of the apparent motion phenomenon suggest the importance of physical matching with single letters. With words, the stimuli are more likely to be coded before the scan and comparisons take place. The lack of a coding stage would thus account for the faster performance on single letters.

TABLE 7
Calculation of Recognition Time

	Alternative delay		
	0 Sec	1 Sec	Difference = recognition time
Words	1198	881	317
Letters	905	807	98

Note: Zero second latencies were measured from the onset of the stimulus display, the one second latencies from the onset of the alternatives. All latencies were corrected for guessing.

Subjective Reports

In answer to a question about the strategies they used to improve performance, only a few subjects failed to report things that they were doing to help identify the stimulus. More complex strategies included rehearsing the stimulus on the longer delay trials, trying to avoid looking at the choices until the stimulus had been identified, and trying to take in the whole field, as one subject had learned to do in a speed reading course.

The strategies reported did not attempt to take advantage of the differences among the conditions. Perhaps this is because only a couple of subjects figured out correctly the shifts of words or letters relative to the fixation point. Most were not at all aware of the relative positions.

About half the subjects thought that they did better on word items than on letters, and a quarter of them thought they did better on letters. Of the three subjects who actually did better on letter items, two thought they did better on words and one didn't know. A few subjects had difficulty answering which item type they did better on because at the exposure durations used they were not able to tell the letter and word items apart.

DISCUSSION

All five of the initial hypotheses have been rejected as complete explanations of the word superiority effect. Performance on words was consistently better than on single letters in all cases, despite the controls suggested by the five hypotheses. It seems appropriate to stop trying to explain away the phenomenon and, instead, to consider the implications for models of the human recognition system.

The major conclusion to be drawn from the strength and persistence of the word superiority effect, as shown in Reicher's (1968) experiment and the experiment reported here, is that word recognition cannot be analyzed into a set of independent letter recognition processes. There is an interaction among the letters such that the context of the other letters of a meaningful word improves recognition despite the control of letter redundancy. It is not a general effect from the context of other letters; Reicher showed that the context of a nonsense quadrigram did not improve performance.

The Serial Versus Parallel Issue

The results of this experiment are not directly relevant to the serial versus parallel issue. The latency data are consistent with either type of model. It is difficult to imagine how a serial model could account for improved performance on four-letter words since the four letters would

take longer to be read off the rapidly fading icon. But such a model cannot be ruled out.

The parallel-serial issue can be restated in terms of the size of the perceptual recognition unit. At some level the aspects of the visual stimulus are processed simultaneously as a unit (i.e., in parallel). This parallel level might be at the stage where individual points are analyzed into lines. In most serial models for words or letter arrays, the processing of the letter features is assumed to be in parallel, but the letters are processed in serial. In even the most radical parallel model, a level is reached at which processing is serial. No one proposes that we read a whole paragraph in parallel. Thus there is not really a dichotomy between serial and parallel processing models. Each model should be specified by the largest unit which is processed in parallel, i.e., at the same time. This unit might be considered the perceptual unit. Possible unit sizes for verbal materials include the letter, spelling-unit, syllable, word, and phrase.

A perceptual unit size can also be based on nonindependence of the sort demonstrated in this experiment. The perceptual unit can be considered the highest level in which facilitative interactions of the lower level units occur in both directions. The mutual facilitation is assumed to occur only when the lower units are processed together as a perceptual unit. For example, an experiment on the perception of word pairs like *TOP HAT* might show that the presence of the first word enhanced recognition of the second, but not vice versa. This would suggest that the first word is processed separately, prior to the second. Thus the perceptual unit would have to be smaller than the word pair. If the results showed mutual facilitation, a perceptual unit size of word pairs or larger would be indicated. Of course, redundancy and set effects would have to be eliminated.

The word superiority results can be interpreted as demonstrating that the perceptual units are larger than single letters. The single letters interact to facilitate performance on words. The perceptual units might be words, but the data do not require this. Any unit size of digram or spelling pattern or larger would account for the word superiority effect. Whether the perceptual unit defined in these two ways coincide remains to be seen.

Feature Analysis

Many models of letter recognition have assumed that the recognition process is based on the extraction of the distinctive features of the stimulus. The extracted features are then used by some decision process to categorize the stimulus. In Selfridge's (1966) "Pandemonium" the lowest level demons are the feature extractors. The operators of Uhr and Vossler's (1963) model extract features from the pattern being recog-

nized. Rumelhart's (in press) components in his multicomponent theory are equivalent to features. The attributes input to the logogen in Morton's (1969) model can also be considered as features.

The feature framework treats individual letters as bundles of features. The features are usually identified with visually apparent attributes of the letters, such as "curved shape on top." Words consist of a sequence of letter bundles of features. The pattern recognition system operates by extracting features from the stimulus and applying a decision procedure to determine the identity of the stimulus. If processing time were available, sufficient features would be extracted for the system to identify the stimulus with very little probability of error. In a tachistoscopic recognition situation, not all of the features of each bundle can be extracted. Errors will result when the features needed for a particular decision fail to be extracted.

A simple specific version of a feature model is one that postulates that the extraction of features from all bundles proceeds in parallel such that in a limited exposure a proportion of the features in each bundle would be extracted. Thus when the stimulus *LOVE* is presented, or when *L* is presented alone, the bundle for the *L* should contain the same proportion of features. Performance should be the same for both cases. Yet the experimental results show that the decision between the alternatives *D* and *L* is more likely to be correct when the stimulus was the whole word *LOVE,* than when it was the single letter *L*. The simple model must be modified to account for these results.

More Features

Perhaps more features relevant to the *D* versus *L* decision can be extracted from the word stimulus than from the single letter. The problem then becomes one of identifying the source of the additional features. It is unlikely that increasing the number of letters in the stimulus increases the rate at which features are extracted from each bundle. The attention assumptions so important in Rumelhart's (in press) multicomponent model would predict just the opposite. As the number of letters in the stimulus array increases, the amount of attention or feature extracting capacity available for each bundle should decrease. In addition to being implausible, the extraction of more features per bundle with multiletter stimuli would predict better performance on quadrigrams than on single letters unless other changes were also made in the model. Reicher (1968) showed that performance on quadrigrams was not better than on letters.

The experiment reported here eliminated two hypotheses that otherwise would have been considered as the source of the additional features.

These are the focusing and response bias hypotheses. The focusing hypothesis supposes that the additional features are extracted at the tested position at the expense of not extracting as many features from other letter positions. The response bias hypothesis suggests that the feature bundles for the untested letters of the word stimuli, i.e., *OVE*, are, in fact, relevant to the decision between the *D* and *L* alternatives. Both of these hypotheses were eliminated by the use of carefully balanced stimulus sets.

There are additional possibilities for the source of more features relevant to the choice between the response alternatives if the restriction of features to letter bundles is relaxed. With multiletter stimuli there could be additional features extracted from the various combinations of letters, independent of the specific letter features. Any feature extracted from a letter combination including the tested letter might be relevant to the choice between the two alternatives. The additional information available from these features would enable the subject to perform better on words than on letters.

The letter combinations from which these additional features are extracted might be the whole word units. Words certainly do have distinctive overall shapes, especially in the pattern of letter heights in lower case form. The word *love*, for instance, has a tall-short-short-short pattern. Overall shape is not as likely a basis for relevant features when the words are written in upper case type, as they were in the present experiment. The patterns of height variations disappear when the words are written in upper case type. Also, most of our experience with words is with lower case type. It is less likely that the decision mechanism in the pattern recognition system is adapted to the use of features from whole words written in upper case type.

Features based on digrams or trigrams are also plausible. The overall shape of pairs is a likely basis for features. For instance, the digram *CO* is rounded overall while *NI* is square. Another possible source of features specific to digrams is the space between the letters. A number of pairs, like *BY*, have fairly distinctive shapes between the letters.

The hypothesis of letter pair features without additional assumptions would seem to predict that quadrigrams, with as many letter pairs as words, would produce performance equal to that of words. Reicher showed that to be false. An assumption that would account for these results is that there is a digram frequency effect. There are a number of possible models for the decision process based on features that would account for better performance with a high frequency digram like *TH* than with the much lower frequency digram *HT*.

Feature Selection

The extraction of more features in the case of word stimuli is not the only way to account for superior performance on words. Of the features extracted from the feature bundle for the letter *L*, only a few may be relevant to the decision between the alternatives *L* and *D*. The others are features on which both *L* and *D* have the same value. Instead of extracting more features in order to get more relevant features, the extraction process could be selective with word stimuli such that of the features extracted for a given exposure time, more will be relevant to the choice between the alternatives. In order for this to explain the superiority of the word stimuli, the features extracted from the irrelevant letters *OVE* must direct the extraction of features from the *L* such that the features which distinguish *LOVE* from other words ending in *OVE* are more likely to be extracted. Since the superiority of words is evident at all letter positions at once, each letter must simultaneously, before it is identified, both affect and be affected by the features extracted from other letter positions.

Discrimination net models, such as EPAM (Feigenbaum, 1963), suggest one type of mechanism that would have the selective properties required for this explanation. The discrimination net has stored at each decision node the name of the feature to be tested. If features are not extracted until required for a test, the system would have the required selective property. The feature extracted at any moment would depend on the test in the current node. The current node, of course, depends on *all* the features extracted previously. The arrangement of the features in the decision nodes determined the efficiency of the use of the information in the extracted features. It should not be difficult to find an arrangement of a discrimination net that would take advantage of the redundancies of English words.

This sketch of a discrimination net model, although based on EPAM, must differ from it in the nature of the factors limiting performance on the model. In EPAM, the limiting factor is usually assumed to be the time required for each test at a node. Thus the processing rate is determined by the number of choices at test nodes that can be made per unit time. This assumption suggests that the overall number of features tested should be the same for both word and letter stimuli. Since word stimuli contain more information, performance should be better on the single letters.

The modification required in order to make EPAM a selective feature extraction model for the recognition of word stimuli is to make the limiting process the extraction of features rather than the decision processes

at the test nodes. Furthermore, this limitation on the rate of feature extraction must apply independently for each position so that four times as many features are extracted from four-letter words as from single letters. The interface between the spatially parallel feature extractors and the essentially serial decision net may be difficult to work out.

Verbal Coding with Information Loss

The preceding two models have attempted to account for superior performance on word stimuli by postulating a source for additional information in the case of word stimuli. The same information difference can be obtained by proposing that information is lost in the case of single letter stimuli. This information loss could occur in the process of categorizing or producing a verbal code for either the word or letter stimulus. All we need to suppose is that the system has only a single verbal code, either a word name or a letter name, available at the point the decision is made between the two forced-choice alternatives.

An example will make this alternative clear. Suppose the single letter *L* is presented and sufficient features are extracted to limit the possible letters that it could be to the set *B*, *E*, *M*, and *L*. If this information is available at the time of a forced-choice decision between *D* and *L*, the correct choice will be made. But if the system must code the feature information into a verbal code before the forced-choice decision, the system would lose the information needed for that decision except when *L* happened to be the code selected.

Now suppose that the stimulus *LOVE* is presented and that the same features are extracted from the *L* as before. The possible letters for the first position of the word are the same set as before, *B*, *E*, *M*, and *L*. But the system now has some basis for selecting among these letters. It looks for a word code which simultaneously satisfies the constraints provided by all four letter positions. The actual stimulus word, *LOVE*, will, of course, satisfy all the constraints.

There may be other words that satisfy all the constraints provided by the features extracted from all letter positions. With *M* as a possibility for the first letter position, *MOVE* is another solution to the simultaneous constraints. *LONE* and *LOSE* are solutions if *N* and *S* are possibilities for the third letter position. Although there may be several solutions, there is likely to be in each letter position some possible letters that do not enter into any of the solutions. In the first position of the example, the letters *B* and *E* may not appear in any of the solutions. When a word is selected from the set of solutions to the simultaneous constraints to be the verbal code for that item, it is more likely to include the letter *L* than in the case where a letter was selected from the set of four possible letters

to be the verbal code for the single letter stimulus. Thus the word code is more likely to retain the information from the feature analysis that allows a correct choice to be made between the alternatives *D* and *L*.

The simultaneous constraints model is almost identical to the fragment theory of Newbigging (1961) and the sophisticated guessing model of Broadbent (1967). But the general notion of information loss in coding is also consistent with the signal detection models of Broadbent (1967) and Morton (1968, 1969).

An additional attraction of the simultaneous constraints model is that it is easy to qualitatively account for the increase in processing time as word length increases. On a more abstract level, the model consists of two stages. The first is a feature extraction stage in which the processing is, presumably, in parallel. The second stage uses the features to find, construct, or otherwise determine a code for the stimulus. Many reasonable models of this coding process predict that the processing time should increase with the length of the stimulus words. The next step is to find a model which matches the negatively accelerated increase in time as a function of word length found by Stewart, James, and Gough (1969).

Conclusions

The three models account for superior performance on words by postulating (*a*) more features from digrams or larger units, (*b*) selection of features for greater efficiency, and (*c*) information loss in verbal coding. Experimental tests of the models are difficult to make. The first two could be distinguished if the features could be identified. Eleanor Gibson has made some progress with confusion matrix methods of identifying the distinctive features used in the recognition of single letters (see Gibson, Schapiro, and Yonas, unpublished), but the techniques are not sufficient for a test of the models. The verbal coding model makes one easily testable prediction.[4] Performance on the forced-choice task with letter stimuli, when corrected for guessing, should be no better than performance when the subject reports a single letter without having alternatives to choose from. This test, of course, would not give direct evidence about the word processing.

REFERENCES

AVERBACH, E., & CORIELL, A. S. Short term memory in vision. *Bell System Technical Journal*, 1961, **40**, 309–328.

BROADBENT, D. E. Word-frequency effect and response bias, *Psychological Review*, 1967, **74**, 1–15.

[4] Pointed out by Gordon H. Bower, personal communication.

CONRAD, R. Acoustic confusions in immediate memory. *British Journal of Psychology,* 1964, **55,** 75–84.

ESTES, W. K., & TAYLOR, H. A. A detection method and probabilistic models for assessing information processing from brief visual displays. *Proceedings of the National Academy of Sciences,* 1964, **52**(2), 446–454.

ESTES, W. K., & TAYLOR, H. A. Visual detection in relation to display size and redundancy of critical elements. *Perception and Psychophysics,* 1966, **1,** 91–16.

FEIGENBAUM, E. A. The simulation of verbal learning behavior. In E. A. Feigenbaum and J. Feldman (Eds.), *Computers and Thought.* New York: McGraw-Hill, 1963.

GIBSON, E. J., SCHAPIRO, F., & YONAS, A. Confusion matrices for graphic patterns obtained with a latency measure. Unpublished manuscript, Cornell University.

HOWES, D. H., & SOLOMON, R. L. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology,* 1951, **41,** 401–410.

KINCAID, W. M. The combination of 2 X *m* contingency tables. *Biometrics,* 1962, **18,** 224–228.

MORTON, J. A retest of the response-bias explanation of the word frequency effect. *British Journal of Mathematical and Statistical Psychology,* 1968, **21,** 21–33.

MORTON, J. Interaction of information in word recognition. *Psycholohy Review,* 1969, **76,** 165–178.

NEISSER, U. *Cognitive psychology.* New York: Appleton-Century-Crofts, 1967.

NEWBIGGING, P. L. The perceptual redintegration of frequent and infrequent words. *Canadian Journal of Psychology,* 1961, **15,** 123–132.

REICHER, G. M. Perceptual recognition as a function of meaningfulness of stimulus material. Technical Report No. 7, 1968, The University of Michigan, Human Performance Center.

RUMELHART, D. E. A multicomponent theory of the perception of briefly exposed visual displays. *Journal of Mathematical Psychology* (in press).

SELFRIDGE, O. G. Pandemonium: A paradigm for learning. In L. Uhr (Ed.), *Pattern recognition.* New York: Wiley, 1966.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

SPERLING, G. The information available in brief visual presentations. *Psychological Monographs,* 1960, **74**(11).

STEWART, M. L., JAMES, C. T., & GOUGH, P. B. Word recognition latency as a function of word length. Paper presented at Midwestern Psychological Association Convention, May, 1969.

THORNDIKE, E. L. & LORGE, I. *The teacher's word book of 30,000 words.* New York: Teachers College Press, 1944.

UHR, L. Pattern recognition. In L. Uhr (Ed.), *Pattern recognition.* New York: Wiley, 1966.

UHR, L., & VOSSLER, C. A pattern-recognition program that generates, evaluates and adjusts its own operators. In E. A. Feigenbaum and J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill, 1963.

WOLFORD, G. L., WESSEL, D. L., & ESTES, W. K. Further evidence concerning scanning and sampling assumptions of visual detection models. Technical Report No. 126, 1968, Stanford University, Institute for Mathematical Studies in the Social Sciences.

(Accepted October 16, 1969)