COGNITIVE PSYCHOLOGY 6, 84–107 (1974)

The Control of Eye Fixation by the Meaning of Spoken Language

A New Methodology for the Real-Time Investigation of Speech Perception, Memory, and Language Processing

ROGER M. COOPER[1,2]
Stanford University

When people are simultaneously presented with spoken language and a visual field containing elements semantically related to the informative items of speech, they tend to spontaneously direct their line of sight to those elements which are most closely related to the meaning of the language currently heard (e.g. fixating a lion upon hearing part or all of the word lion, and then a lion, zebra, and snake upon hearing the word Africa). Such behavior may be viewed as an active on-line anticipative process, whereby the contemporary visual field of an observer is subjected to his continual interpretation in terms of the language heard; and, conversely, as a process in which continuous speech is interpreted from moment to moment in the context of the contemporary visual field. In the present study, approximately 55% of all appropriately directed fixation responses elicited by the informative words of a prose passage were initiated even while these words were being pronounced (in a number of instances on the initial phonemes), and nearly 40% of post-word responses were found to occur within the first fifth of a second following word termination. The linguistic sensitivity of this response system together with its associated small latencies suggests its use as a practical new research tool for the real-time investigation of perceptual and cognitive processes and, in particular, for the detailed study of speech perception, memory, and language processing.

Although recent years have witnessed the accumulation of a vast body of literature on eye movement recording and research (e.g. Fender, 1964; Mackworth, 1968; Mackworth & Morandi, 1967; Noton & Stark, 1971;

[1] Currently with the Department of Psychology, University of Oregon and Stanford Research Institute. Requests for reprints should be sent to Roger M. Cooper, Department of Psychology, University of Oregon, Eugene, Oregon, 97403.

[2] This work was done during 1972–1973 while the author was associated in postdoctoral research with Norman H. Mackworth in the Department of Psychology at Stanford University. The author is greatly indebted to Norman Mackworth for his invaluable encouragement and support and for the use of his eye movement camera. The author also wishes to thank Richard C. Atkinson and Jane F. Mackworth for their help.

Yarbus, 1967; Young, 1963), no attention has previously been given to (1) the investigation of any possible relationship between the locus of eye fixation and the meaning of concurrently heard or spoken language, and (2) the possibility of exploiting such a relationship (assuming that one exists) as a research tool for the detailed investigation of perceptual and cognitive processes.

That language in some form is capable of exercising at least a small degree of control over eye movements has recently been demonstrated by Carpenter and Just (1972), who showed that the interpretation given to a previously read ambiguous sentence could be inferred from the subsequent visual selection of an appropriate target. Additionally, Williams (1966) showed that the gaze selects objects in concordance with a previous written specification of attributes.

The present paper describes a study designed to investigate the extent to which the meaning of spoken language is able to control the locus of eye fixation, and proposes the use of the resulting technique as a practical research tool having important advantages over currently available techniques. For example, if it were indeed true that the meaning of ongoing heard language in conjunction with the contemporary visual field played a significant role in determining where one looks or does not look from moment to moment, then the specification of the nature and order of targets fixated, together with the associated latencies and fixation times (time spent on target), would appear to provide sensitive measures for determining how people interpret and process spoken language from moment to moment without interrupting the continuity of spoken language presentation. If, in addition, the latencies involved were sufficiently small, this could also have the advantage of eliminating the presence of sudden attentional distractions and short-term memory losses inherent in the application of traditional gross motor movement or relatively long latency response measures, such as key pressing and verbal reporting.

The main objective of this study was to test the following hypothesis: *When people are simultaneously presented with spoken language and a visual field containing elements semantically related to the informative items of speech, they tend to spontaneously direct their line of sight to those elements which are most closely related to the meaning of the language currently heard.* In particular, the scanning of those elements which are most closely related to a given word should be initiated during or closely following the pronunciation of that word, and without prior instruction to do so. Such behavior, if confirmed, could be viewed as an active on-line anticipative process whereby the contemporary visual field of an observer is subjected to his continual interpretation in terms of the language heard; and conversely, as a process in which ongoing heard

language is interpreted word by word in the context of the contemporary visual field.

METHOD

Subjects and Design

Forty male and female Stanford students ranging in age from 18 to 30 were randomly assigned to two independent groups of equal size with repeated measures on word–picture relation category and type of instruction. The Ss selected were volunteers required to have normal vision and hearing, as well as fluency and command of the English language.

Stimulus Materials

Visual stimulus materials consisted of four slides (see Fig. 1), each of which was composed of black and white line drawings of nine distinct commonplace objects (e.g. queen, lion, barn, sailboat) with white surrounds and arranged in the form of a 3×3 matrix, subtending an angle of 30° along the diagonal when viewed by S. All drawings appeared superimposed on a black grid composed of nine square cells of equal size, with the central position of each cell occupied by a different stimulus target. The uniform arrangement of target elements into 3×3 arrays of equal dimensions was used to simplify subsequent scoring and calibration procedures.

Auditory stimulus materials consisted of tape-recordings of four prose passages of durations 2 min 36 sec, 1 min 33 sec, 1 min 37 sec, and 1 min 47 sec, respectively, and two comprehension tests of durations 2 min 16 sec and 2 min 10 sec, respectively, narrated by *E* and presented to S at a rate of 145 words/min. The prose passages were fictional short stories constructed so that a large proportion of highly informative words and compound expressions (e.g. king of the beasts), decided beforehand as critical items to be scored, would make either direct or indirect reference to specified subsets of pictures on a corresponding slide. For example, in Story 3, the words *lion* and *zebra* referred directly to individual pictures of a lion and a zebra, while the word *Africa* referred indirectly to individual pictures of a lion, a zebra, and a snake. Conversely, each picture referred to a set of words within the corresponding prose passage. For example, in Story 1, the words *queen, agony, she, cross, flustered, herself*, etc., referred to a single picture of a queen. In this manner there was defined a multi-valued relation from the informative words of speech within a prose passage to the set of all pictures on a corresponding slide. This relation was primarily single-valued (i.e., one picture associated with each word) in the cases of direct noncontextual and direct context-

ual relations defined below. Each comprehension test consisted of 15 questions (10 of which were sentence completion type), with each question typically of duration 2–3 sec, and followed by a 5 sec answer period. Both tests were constructed so that the answers to all questions were represented pictorially (in most cases as unique target elements) on the slide exposed to S.

Fundamental Definitions

Word–picture relation categories. The referential relations between the critical items of a prose passage and the pictures of a corresponding slide were divided into four mutually exclusive and exhaustive categories, with classification proceeding according to (1) whether a critical item referred directly or indirectly to at least one picture of the corresponding slide, and (2) whether relations established in this manner took into account previous verbal context. Relations were classified as follows:

(1) *Direct Noncontextual* (involving primarily nouns and adjectives), if they incorporated the largest set of pictures (usually one picture), each of which was (or contained) an exact representation of the corresponding pronounced word, when this word was interpreted in isolation from the previous verbal context. Differences of plurality and personal identification between critical items and the pictures to which they referred were ignored. For example, the narrator's *dog* in Story 3 and the word *sailboats* mentioned in Story 4 were considered to be directly and noncontextually related to pictures of an anonymous dog and a single sailboat, respectively.

(2) *Direct Contextual* (involving primarily nouns and pronouns), if they incorporated the largest set of pictures (usually one picture), each of which was (or contained) an exact representation of the corresponding pronounced word after taking into account the previous verbal context. Again, differences of plurality and personal identification were ignored. It was usually (but not always) the case that in the process of considering previous verbal context, the number of such pictures was reduced (i.e., became a proper subset of the corresponding noncontextual interpretation). For example, although the words *she* and *herself* were directly and noncontextually related to each of the animals pictured on a corresponding slide (a five element subset, including a queen), they were contextually related only to the picture of the queen, when taking into account their antecedent *queen*, appearing in the previous sentence.

(3) *Indirect Noncontextual* (involving primarily nouns, adjectives, transitive verbs, adverbs), if they incorporated the largest set of pictures, each of which was closely, but not directly, related to the pronounced word (or compound expression) by means of readily apparent associa-

tions, when this word or expression was interpreted in isolation from the surrounding verbal context (e.g. the word *lake* corresponding to a picture of a sailboat and the word *king* corresponding to pictures of a queen and a lion (king of the beasts)).

(4) *Indirect Contextual* (involving primarily nouns and pronouns), if they incorporated the largest set of pictures, each of which was closely (but not directly) related to the pronounced word (or compound expression) by means of readily apparent associations after taking into account the previous verbal context. For example, in the sentence "The queen was in agony.", the word *agony* was indirectly and contextually related to the picture of a queen; and in the phrase ". . . cooling herself with a peacock feather fan until the handle snapped in two.", the word *handle* was indirectly and contextually related to the picture of a peacock.

In order to insure mutual exclusion of categories, all ambiguities of classification were resolved in favor of the direct relation and of relations indicating comprehension of verbal context (i.e., critical items which were involved in both direct and indirect word–picture relations had only their direct relations scored, and those involved in both contextual and noncontextual relations had only their contextual relations scored). Intransitive verbs (e.g. is, was, are), articles (e.g. a, an, the), conjunctions (e.g. and), disjunctions (e.g. or), and prepositions (e.g. of, from, because) were regarded as relatively uninformative and, hence, not considered to be critical items. Indirect relations which were not self-evident and whose existence was a matter of personal judgment (e.g. the word *Africa* and a picture of a peacock, or the word *lion* and a picture of a zebra) were also ignored.

Definition of correct responding. The subject was understood to execute a correct (α,β)-fixation response to a critical item of speech and, hence, to verify the corresponding word–picture relation, if either of the following two conditions was satisfied: (1) Fixation of the associated target set (i.e., the scanning of any subset of its elements) was initiated at any time during the pronunciation of a word and persisted at least through word termination (within-word response). (2) Fixation of the associated target set was initiated within α sec following word termination, or else prior to commencement of the next word having the same target set as referent, whichever came first, and such fixation persisted for at least β sec (post-word response).

In the present study, the α and β criterion levels were chosen to be 1 sec and $\frac{1}{3}$ sec, respectively. Furthermore, it was understood that condition (2) was applicable, only if S failed to satisfy condition (1), and that if S executed more than one correct post-word response to a given critical item, only the first would be counted. This choice of definition

and parameter values was motivated by (a) the desire to obtain an optimal definition of correct responding, based on a priori considerations of the kinds of responses Ss most frequently tend to give, and (b) a criterion of maximizing built-in reliability by reducing to a minimum the possibility of counting S's miscalculations as correct responses. Fixations directed to the associated target sets occurring prior to the pronunciation of a word were not counted as correct responses, even if such fixations persisted through word termination. This was due to possible confounding with behavior sometimes arising in which S was observed to continue fixating the same point independent of the meaning of concurrently heard language. Thus, for S to execute a correct response, his direction of eye fixation had to be appropriately altered either during or closely following the pronunciation of a word.

Categories of correct responding. Correct fixation responses were identified with the aforementioned word–picture relations (i.e., direct noncontextual, etc.). Within-word responses were further subdivided into the following three catgories: (1) Initial responses (I-responses)—correct responses initiated at the beginnings of pronounced words (i.e., during the pronunciation of the smallest subset of initial letter sounds which could be distinguished from the sounds of letters that followed). For example, in each of the following words the beginning of the arrow indicates the point within the pronounced word at which fixation of the associated target set is initiated, and the length of the arrow indicates the duration of the fixation relative to the pronounced word (zebra, queen, she, sailboats). (2) Terminal responses (T-responses)—correct responses initiated at the ends of pronounced words (i.e., during the pronunciation of the smallest subset of terminal letter sounds which could be distinguished from the sounds of letters that preceded (e.g. lion, zebra, queen, king, umbrella)). (3) Between responses (B-responses)—correct responses commencing later than initial responses but prior to terminal responses.

The same categories were also applied to the investigation of withdrawal responses (rapid looking away behavior elicited by pronouncing the name (frequently the initial phonemes) or other strongly associated verbal referent of an item whose fixation is already in progress). In this case the beginnings of the arrows indicate where within the pronounced words looking away behavior is initiated.

Not all critical items had a between section (e.g. she) and ambiguities in classification were resolved in favor of the temporally earlier interpretation. For example, a within-word response to *he* was considered to be an I-response rather than a T-response. Post-word responses were classi-

fied dichotomously according to whether or not correct responding was initiated during the first fifth of a second following word termination.

Performance on Comprehension Tests

Performance on both comprehension tests was based upon the execution of correct eye fixation responses to the anticipatory cue words of each question. A word within a question was considered to be an anticipatory cue word if it, together with the previously heard words of that question, potentially caused a reduction in size of the set of all pictures which were candidates for the correct answer, by increasing the number of associations with that picture which represented the correct answer. For example, the first question of Comprehension Test 2 was "In the *beginning* of the *story* what was *George/doing*?", and the italicized words were considered to provide anticipatory cues to the correct answer, *typing,* represented by a picture of a typewriter. The subject was instructed to verbally report the answer to each question as soon as possible after hearing the first word. If S initiated a correct verbal report to a question prior to its termination or within the following 5-sec answer period, performance was based upon (a) whether or not S correctly fixated the answer (e.g. the typewriter in the previous example) at any time during the statement of the question or within the allotted 5-sec answer period, but prior to the initiation of his correct verbal report (total or T-responses), (b) whether or not visual selection of the correct answer occurred prior to question termination and the initiation of a correct verbal report (within-question or WQ-responses), and (c) whether or not visual selection of the correct answer occurred during the pronunciation of the first anticipatory cue word (earliest possible or EP-responses). (The reader should note that in the application of the definition of correct responding to comprehension test performance, post-word responses were required to be initiated prior to the next consecutive anticipatory cue word.) In those cases where S gave either an incorrect or no verbal report within the allotted 5-sec answer period, performance categories for the visual selection of correct answers were defined similarly, except that the constraint of responding prior to the initiation of a correct verbal report was removed. For the sake of completeness, information was desired on the frequency and latency of fixation responses directed to targets which were pictorial representations of subsequent incorrect verbal reports (e.g. fixating lion and then verbally reporting *lion* when the correct answer is *zebra*).

Apparatus and Equipment

All eye movements were monitored by the corneal reflection technique using the most recent version of the Mackworth eye movement camera

system (adapted from Mackworth, 1968), capable of an accuracy of $\pm 1°$ in a $20° \times 20°$ field. Throughout the experimental session S's head remained fixed in one position through the use of chin and forehead rests. The slides, which were rear projected on an 8 in. × 8 in. ground glass screen located 15 in. away from and facing S, were of sufficiently high contrast to allow their reflections to be clearly observed on the cornea of S. A television camera with 200 mm telephoto lens was trained on S's right eye from a distance of 20 in., at approximately the same height, and 30° to the right of S's mean line of sight. An enlarged version of S's right eye with the scene being viewed superimposed on S's pupil could be observed on a nearby television screen not visible to S. In this manner all eye movements were subject to on-line, real-time observation without S's knowledge. A video recorder equipped with sound track and slow motion control provided a simultaneous record of all audio–visual materials together with S's sequence of eye fixations. This allowed a simple and direct method of correlating what S heard with where he looked in subsequent scoring procedures. The slow motion control was used to gather data on response latencies and fixation times as integral multiples of $\frac{1}{60}$ sec, the time taken for the cathode ray beam to make one vertical sweep across the screen.

Procedure

At the beginning of each session S was (1) misinformed as to the true purpose of the experiment by telling him that its purpose was to monitor changes in his pupils while he was simultaneously viewing slides and listening to prose passages, and (2) told that he could look anywhere he liked on the screen in front of him, but that at no time was he to take his eyes off the display other than to blink. The subject was then presented with the following materials in the order specified: (1) Stories 1, 2, and 3, each accompanied throughout its duration by its corresponding relevant slide and a 3-sec period of prior visual exposure, (2) a surprise auditorily presented comprehension test on Story 3 with continued exposure to the same slide, (3) an instruction to S that he was to listen as carefully as he could to the following story, as he would be tested for comprehension upon its completion, (4) Story 4 accompanied throughout by its corresponding relevant slide and a 3-sec period of prior visual exposure, and (5) the expected auditorily presented comprehension test on Story 4 with continued exposure to the same slide. The session terminated with calibration of S's eye movements followed by S's post-experimental verbal report.

Control Ss received exactly the same instructions, prose passages, and comprehension tests, in the same order. The accompanying slides were

also the same, but were presented in a different order so that S now saw slides largely irrelevant to the stories he was listening to. The previously given definition of correct responding was also applied to the scoring of control S responses, except that in this case a response was considered to be correct, if S appropriately fixated the elements of those cells (now containing irrelevant pictures) which for the experimental group contained members of the associated target set. If the direction of eye fixation were independent of the meaning of spoken language (null hypothesis), one would not expect there to be a significant difference between the number of correct fixation responses to the target elements of a relevant slide and those directed to irrelevant targets occupying the same cell positions when identical critical item sets are employed in both cases.

Illustration of Fixation Performance

The following illustrates typical performance to Story 3 and Comprehension Test 1. All critical items of Story 3 and anticipatory cue words of Comprehension Test 1 are italicized, with the arrows immediately above each line indicating the location and duration of eye fixations relative to the language heard. The abbreviation symbols (Z, D, S, C, L, T, Pk, P, G) have been taken to represent the *zebra, dog, snake, camera, lion, tree, peacock, pipe,* and *grapes,* respectively, of the corresponding slide pictured in Fig. 1.

While on a *photographic/safari* in *Africa*, I managed to get a number of breathtaking shots of the wild *terrain*. These included pictures of rugged mountains and *forests*, as well as muddy *streams/winding their way* through *big game/country*. One of my best shots though was ruined by my scatter-brained *dog/Scotty*. Just as I had slowly *wormed* my way on my *stomach* to within range of a *flock* of magnificent *birds,/feasting* on *bunches* of wild *grapes*, entwined in the *branches* of a small *tree, Scotty/bounded up/barking/madly, and the birds/scattered in all directions*. With a sigh I lit my *pipe*, wondering how *he* had gotten away from camp, when suddenly I noticed a hungry *lion* slowly *moving* through the tall grass toward a herd of grazing *zebra*. As I grabbed my *camera*, I could see *he* had singled out a *member* of the *herd*. But before *he* could *pounce*, the *zebra/bolted*, causing all the *animals to stampede*. The *frustrated/lion* looked as *stupified* as I felt,

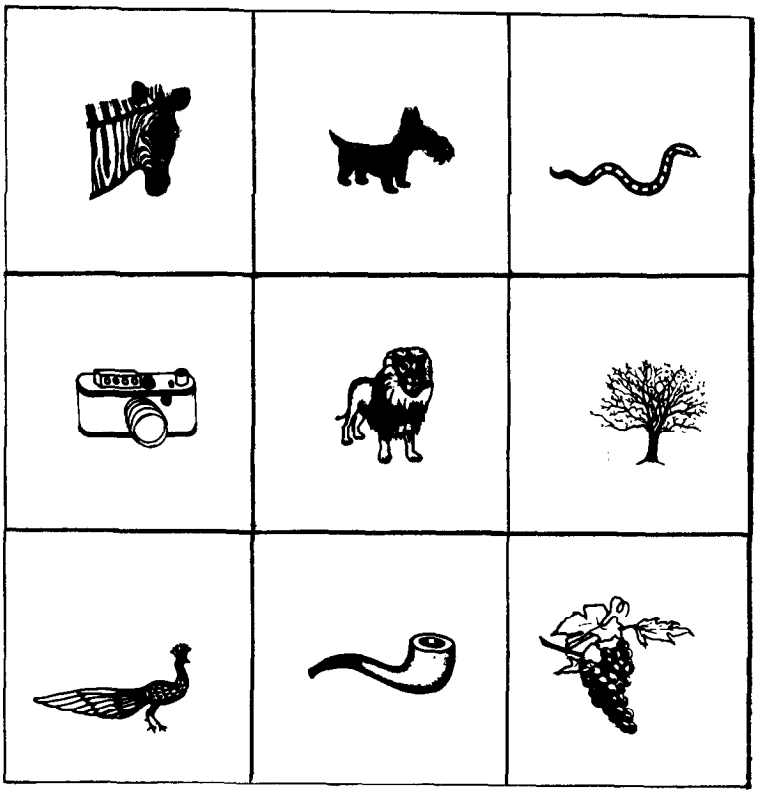

FIG. 1. The slide accompanying Story 3.

Pk——→S——→Z—
until I noticed *slithering* through the grass a long *snake*, just where the *zebra*
→ Z——→ S→ L——→ Z——→
had been *grazing*. The *king of the beasts* hadn't gotten *his/prey*, but I had
D→ D→ D——→L→S→ D
some good pictures, and *Scotty* was so *frightened* by the events that *he*
——→ D——→ L——→Z——→S——→
never again left camp the whole time *we* were in *Africa*.

Critical items and word–picture relations of Story 3 in order of presentation.

(1) Direct Noncontextual Relations (20): forests(T), big game(L,Z), dog(D), wormed(S), birds(Pk), bunches(G), grapes(G), branches(T), tree(T), birds(Pk), pipe(P), lion(L), zebra(Z), camera(C), zebra(Z), animals(L,Z,S,D,Pk), lion(L), snake(S), zebra(Z), king of the beasts(L).

(2) Direct Contextual Relations (12): Scotty(D), Scotty(D), he(D), he(L), member(Z), herd(Z)*, he(L), his(L), prey(Z), Scotty(D), he(D), we(D).
(3) Indirect Noncontextual Relations (14): photographic(C), safari(L, Z,S), Africa(L,Z,S), terrain(T,G), streams(S), winding their way(S), country(T,G), flock(G), feasting(G)*, barking(D), scattered in all directions(G,T)*, herd(G)*, slithering(S), Africa (L,Z,S).
(4) Indirect Contextual Relations (13): stomach(S), feasting(Pk)*, bounded up(D), madly(D), scattered in all directions(Pk)*, moving(L), pounce(L), bolted(Z), stampede(L,Z), frustrated(L), stupified(L), grazing(Z,S), frightened(D).

Note that the starred items were ambiguously classified so as to study S's real-time resolution of ambiguities.

Comprehension Test 1 (partial listing)

Instruction. "You are to answer the following questions out loud as quickly as you can. In doing so, try to interrupt the statement of each question as soon as possible after hearing the first word."

Question	S's verbal reply	Correct answer
(1) The photographic safari was *in* . . .	"Africa"	Africa
(5) One of his best *shots* was *ruined/by* . . .	—	his dog Scotty
(10) After he lit his *pipe* he *noticed* . . .	"a lion"	a lion
(11) The lion was going *to/attack/a* . . .	"zebra"	zebra
(13) In the *grass* was a . . .	"snake"	snake
(14) Where in the *grass* was the *snake/located*?	"near the lion"	where the zebra had been
(15) In addition to the zebra, who was *frightened*?	"Scotty"	his dog Scotty

Note that questions were never interrupted with correct verbal reports.

TABLE 1
Mean Percent Correct Fixation Responses per Word-Picture Relation Category

Word–picture relation category	Experimental group					Control group				
	1	2	3	4	P	1	2	3	4	P
Direct Noncontextual	39.1[a] (125/320)	39.0[a] (156/400)	33.5[a] (134/400)	34.6[a] (90/260)	37.0[a] (414/1120)	9.7 (31/320)	14.5 (58/400)	7.8 (31/400)	6.9 (18/260)	10.7 (120/1120)
Direct Contextual	20.0 (60/300)	22.3[b] (49/220)	26.2[a] (63/240)	—	22.6[a] (172/760)	12.7 (38/300)	13.6 (30/220)	10.0 (24/240)	—	12.1 (92/760)
Indirect Noncontextual	21.6[a] (160/740)	35.8[a] (86/240)	29.3 (82/280)	20.2[a] (85/420)	26.0[a] (328/1260)	9.9 (73/740)	12.5 (30/240)	15.4 (43/280)	5.0 (21/420)	11.6 (146/1260)
Indirect Contextual	18.2[a] (102/560)	24.7[a] (74/300)	26.9 (70/260)	20.7[a] (62/300)	22.0[a] (246/1120)	6.8 (38/560)	11.3 (34/300)	15.8 (41/260)	5.7 (17/300)	10.1 (113/1120)

Note. Story numbers are indicated at tops of columns. P = pooled results for Stories 1, 2, and 3. All denominators represent maximum possible scores. Story 4 contained no critical items in the direct contextual category.

Unstarred comparisons are significant ($p < .05$).

[a] Highly significant ($p < .001$).

[b] Not significant ($p = .08$).

RESULTS

The main hypothesis was supported by showing that for each word–picture relation category, the frequency of (α,β)-fixation responses directed to pictures most closely related to the meanings of the words heard was significantly greater than the frequency of those (α,β)-fixation responses directed to the same cells (in the same order) when these cells were filled by irrelevant targets and the same critical item sequence was heard (i.e., the main hypothesis was formulated in terms of a one-tailed criterion). Since the data consisted of frequencies whose underlying distributions were unknown, a nonparametric procedure (one-tailed, Mann–Whitney U-test) was employed in deciding significance. The same technique was also used in establishing results on comprehension.

Frequency Data

The mean percentages of correct fixation responses for all stories and word–picture relation categories are given in Table 1. Corresponding pooled results for Stories 1, 2, and 3 are displayed graphically in Fig. 2. (The results of Story 4 were not pooled with those of Stories 1, 2, and 3, since Story 4 was accompanied by a different instruction.) With the exception of Story 2, direct contextual relations ($U = 147$, $p = .08$), all results were significant ($U < 129$, $p < .05$), and in all but three cases highly significant ($U < 74$, $p < .001$). Significance was obtained for the direct contextual relations category ($U = 1044$, $p < .001$) by pooling results for Stories 1, 2, and 3.

As seen from Fig. 2, the highest frequency of correct responding (37% for Stories 1–3 pooled) was associated with the direct noncon-

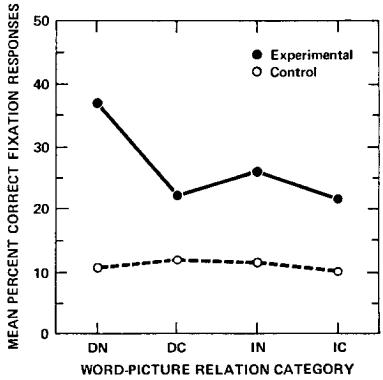

FIG. 2. Mean frequency of correct fixation responses per word–picture relation category.

textual category. This value was significantly greater than corresponding results obtained for the remaining three categories ($p < .001$). The next highest category, indirect noncontextual relations, achieved 26% correct fixation responses, only slightly greater than corresponding percentages in the two contextual categories, and the obtained differences were not significant.

Latency Data

Information on response latencies is provided in Fig. 3 and Tables 2 and 3. Table 2 gives for each story and word–picture relation category the percentage of correct fixation responses which are within-word responses together with the percentage of admissible post-word responses initiated within 0.2 sec following word termination. Inspection of Table 2 shows that (a) approximately 55% of all correct fixation responses were initiated prior to word termination, (b) nearly 40% of admissible post-word responses were initiated within 0.2 sec following word termination, and (c) that the shortest latency post-word responses (as measured by the highest frequency of PW* responses) may be associated with the direct noncontextual category ($p < .05$). Post-word responses initiated at times greater than one second following word termination were not counted as correct responses to the prose passages of this study. In any event, these constituted a relatively small number (fewer than 10%).

Table 3 provides more detailed information on latencies for within-word responses in the case of direct noncontextual relations. Observe that nearly 20% of correct within-word responses were initial responses; whereas, terminal responses account for approximately 60% and, hence, the largest proportion of within-word responses. This, together with the

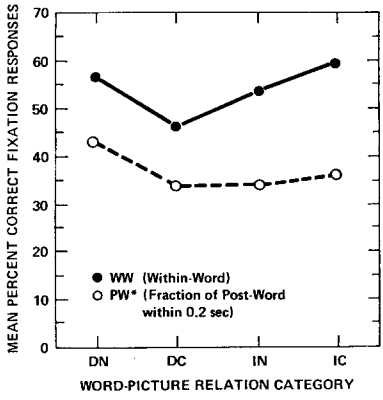

FIG. 3. Partial distribution of response latency per word–picture relation category.

TABLE 2
Partial Distribution of Response Latency per Word–Picture Relation Category

Word–picture relation category	Response category									
	Percent WW					Percent PW*				
	1	2	3	4	P	1	2	3	4	P
Direct Noncontextual	56.0 (70/125)	56.4 (88/156)	59.0 (79/134)	62.0 (56/90)	57.1 (237/415)	47.3 (26/55)	36.8 (25/68)	49.1 (27/55)	38.2 (13/34)	43.3 (77/178)
Direct Contextual	55.0 (33/60)	46.9 (23/49)	41.3 (26/63)	—	47.1 (82/172)	33.3 (9/27)	34.6 (9/26)	35.1 (13/37)	—	34.4 (31/90)
Indirect Noncontextual	60.0 (96/160)	47.7 (41/86)	46.3 (38/82)	54.1 (46/85)	53.4 (175/328)	29.7 (19/64)	31.1 (14/45)	45.5 (20/44)	35.9 (14/39)	34.6 (53/153)
Indirect Contextual	55.9 (57/102)	66.2 (47/74)	58.6 (41/70)	54.8 (34/62)	58.9 (145/246)	40.0 (18/45)	29.6 (8/27)	37.9 (11/29)	46.4 (13/28)	36.6 (37/101)

Note. Percent WW (Within-Word) gives percentage of total correct responses initiated prior to word termination. Percent PW* gives percentage of admissible post-word responses initiated within 0.2 sec following word termination. Story numbers are indicated at tops of columns. P = pooled results for Stories 1, 2, and 3. Story 4 contained no critical items in the direct contextual category.

TABLE 3
Within-Word Distribution of Response Latency: Direct Noncontextual Relations

Story	Response category		
	Percent I	Percent B	Percent T
1	18.6 (13/70)	22.9 (16/70)	58.6 (41/70)
2	15.9 (14/88)	22.7 (20/88)	61.4 (54/88)
3	20.3 (16/79)	22.8 (18/79)	57.0 (45/79)
4	23.2 (13/56)	17.9 (10/56)	58.9 (33/56)
P	18.1 (43/237)	22.8 (54/237)	59.1 (140/237)

Note. I = Initial Responses; B = Between Responses; T = Terminal Responses. Denominators represent total number of within-word responses per story. P = pooled results for Stories 1, 2, and 3.

present data on PW* responses, suggests that the locus of most frequent responding is either situated at the ends of words or within 0.2 sec following their termination. Because the present study invoked in some cases adjacent critical items having the same target element(s) as referent, there is a strong possibility that a certain number of initial responses may in actuality have been short latency post-word responses to immediately prior words. Hence, the actual frequencies of initial responses may not be as high as those given in Table 3, while the actual frequencies of PW* responses are likely to be higher than those reported in Table 2.

Note. In order to increase the sensitivity of the technique to the observation of I-responses, it is recommended that latencies be measured from the beginning of an admissible saccade, rather than from the time at which fixation of the associated target set initiates.

Withdrawal Response Data

An account of withdrawal responding in the case of direct noncontextual relations is provided in Table 4. The percentages given there yield the proportion of prior-to-word responses which became within-word withdrawal responses initiating at the specified locations I, B, or T when critical items were pronounced. A prior-to-word response was one in which the correct target was fixated prior to the pronunciation of a critical item, based upon anticipatory cues extracted from previously heard words such as adjectival modifiers, with the fixation persisting at least through critical item termination. Of special interest due to its application to phoneme perception is the relatively high frequency of I-type withdrawal responses (i.e., those based upon the initial letter sounds of critical items).

TABLE 4
Distribution of Withdrawal Responses:Direct Noncontextual Relations

Story	Response category			
	Percent I	Percent B	Percent T	Percent WW
1	44.4 (4/9)	44.4 (4/9)	11.1 (1/9)	36.0 (9/25)
2	20.0 (3/15)	33.3 (5/15)	46.7 (7/15)	34.1 (15/44)
3	47.1 (8/17)	41.2 (7/17)	11.8 (2/17)	29.8 (17/57)
4	33.3 (3/9)	11.1 (1/9)	55.6 (5/9)	31.0 (9/29)
P	36.6 (15/41)	39.0 (16/41)	24.4 (10/41)	32.5 (41/125)

Note. I = Initial Responses; B = Between Responses; T = Terminal Responses; WW (Within-Word) = Withdrawal Responses occuring prior to word termination (WW = I + B + T). Percent WW gives percentage of prior-to-word responses which became within-word withdrawal responses when critical items were pronounced. P = pooled results for Stories 1, 2, and 3.

Comprehension Data (Surprise Test)

The results of both comprehension tests are summarized in Fig. 4 and Table 5. It is seen that over 60% of the time when experimental Ss executed subsequent correct verbal reports, they executed prior to question termination, correct fixation responses to the anticipatory cue words of questions of the surprise Comprehension Test 1; whereas, control Ss responded in the same manner less than 20% of the time to irrelevant targets occupying cells which for the experimental group contained the correct answers ($U = 12$, $p < .01$). Furthermore, over 45% of the time visual selection of correct answers in the experimental group (as compared with less than 20% for the control group) was based upon hearing

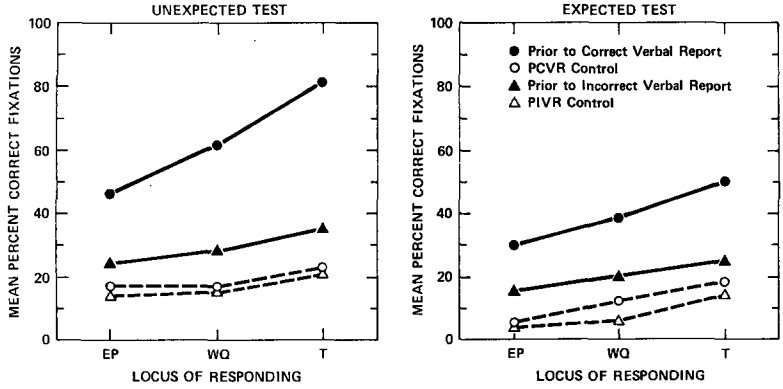

FIG. 4. Comprehension test performance as revealed through fixation responses to anticipatory cue words of questions.

TABLE 5
Comprehension Test Performance as Revealed Through Fixation Responses to Anticipatory Cue Words of Questions

	Comprehension test 1 (unexpected)						Comprehension test 2 (expected)					
	Exper. group 46.0% (69/150)			Control group 34.0% (51/150)			Exper. group 70.7% (106/150)			Control group 68.7% (103/150)		
Response category	EP	WQ	T	EP	WQ	T	EP	WQ	T	EP	WQ	T
Percent correct fixations prior to correct verbal report	46.4 (32/69)	62.3 (43/69)	81.2 (56/69)	17.6 (9/51)	17.6 (9/51)	23.5 (12/51)	30.8 (33/107)	38.3 (41/107)	50.5 (54/107)	5.8 (6/103)	12.6 (13/103)	18.4 (19/103)
Percent correct fixations prior to incorrect verbal report	24.7[a] (20/81)	28.4 (23/81)	34.6 (28/81)	14.1 (14/99)	15.2 (15/99)	21.2 (21/99)	15.9[a] (7/44)	20.5[a] (9/44)	25.0[a] (11/44)	4.3 (2/47)	6.4 (3/47)	14.9 (7/47)
Percent fixations corresponding and prior to incorrect verbal report	11.1[a] (9/81)	16.0[a] (13/81)	21.0[a] (17/81)	7.1 (7/99)	9.1 (9/99)	11.1 (11/99)	2.3[a] (1/44)	2.3[a] (1/44)	2.3[a] (1/44)	2.1 (1/47)	4.3 (2/47)	4.3 (2/47)

Note. Tabulated results are cumulative: EP = Earliest Possible Responses (i.e., responses initiated during first anticipatory cue word); WQ = Within-Question Responses (i.e., responses initiated prior to question termination and which include EP-responses): T = Total Responses (i.e., T = WQ + PQ where PQ = Post-Question Responses). Results given are for *Ss* 1–10 in each group. Listings immediately to the right of group titles indicate percentage of questions verbally answered correctly. Denominators represent total number of questions verbally answered correctly (incorrectly).

[a] Not significant.

only all or part of the first anticipatory cue word and, hence, upon a small, possibly minimal, number of phonological and syntactic cues ($U = 20.5$, $p < .05$). On the other hand, neither group ever initiated execution of a correct verbal report prior to question termination during Comprehension Test 1. It is also evident from Table 5 that during Comprehension Test 1, when Ss of both groups gave incorrect or no verbal reports within the allotted 5-sec answer periods, experimental Ss visually selected the correct answers to questions prior to question termination nearly 30% of the time; whereas, control Ss responded in the same manner approximately 15% of the time ($U = 25$, $p < .05$). These results suggest that visual selection of correct answers as a performance measure for testing comprehension may be more sensitive and associated with a more liberal strategy of responding than traditional verbal reporting of answers. Hence, the use of this technique as a new and improved method of testing comprehension appears promising.

Effect of Instruction to Listen for Comprehension

Table 1 shows that an instruction which directs maximal attention to the auditory modality and simultaneously announces a comprehension test did not improve overall fixation performance to Story 4. In fact, the percentage of Ss executing fewer than 10% correct fixations increased from 5% to 20% in passing from Story 3 to Story 4. As is evident from Table 5, a marked decrease in such performance was clearly observed during the expected Comprehension Test 2 in those cases involving the visual selection of correct answers prior to correct verbal reports ($U < 25$, $p < .05$; WQ and T), although between-group comparisons remained significant ($U < 25$, $p < .05$). This was realized as an acquired tendency to continue fixating the same location independent of the meaning of concurrently heard language. On the other hand, the percentage of questions verbally answered correctly increased from 46.0% to 70.7%, suggesting that this tendency may have resulted from the adoption of a strategy to block informative visual input as a means for improving subsequent verbal performance through the elimination of influences believed to be distracting. That Ss believed this to be the case was evident from their post-experiment verbal reports.

Additional Observations

Throughout the course of the present study three main types of visual behavior were observed: (1) a visual–aural interaction mode, in which fixation of targets was correlated with the meaning of concurrently heard language, (2) a free-scanning mode, in which S continually altered his direction of gaze in a manner independent of the meaning

of concurrently heard language, and (3) a point-fixation mode, in which S continued to fixate the same location independent of the meaning of concurrently heard language. It was frequently the case that Ss would vacillate between more than one of these modes during the presentation of a single story or comprehension test. Informal evidence based upon Ss' post-experimental verbal reports suggests that these three types of visual behaviors may be related to their distribution of attention between the visual and auditory modalities.

It was also observed that Ss often executed fixation responses to words other than those belonging to the established critical item sets, and, in particular, to articles, prepositions, and conjunctions. For example, upon hearing part or all of the word *and* in "There would be lions and," a number of Ss switched their direction of gaze from a lion to another animal such as a zebra. Such behavior once again reveals the sensitive anticipatory characteristic of the response system under investigation, and further indicates that so called "uninformative words" may possess sufficient information content to trigger correlated fixation responses. At other times, Ss appeared to fixate targets based upon their own personal idiosyncratic interpretations of concurrently heard words (i.e., subjective indirect associations), with no observable consistency across Ss.

Additionally, Ss sometimes executed fixation responses to words as though they had interpreted these words in isolation, and without reference to the previous verbal context. For example, in connection with the following excerpt from Story 1, rather than fixating the picture of a peacock, five Ss directed their fixations to a vase with a handle, upon hearing part or all of the word *handle* (even though the contextual referent of *handle* is the peacock feather fan, associated with the peacock); and five other Ss directed their fixation to a zebra upon hearing part or all of the word striped (even though *striped* is referentially related to the king's forehead).

> All day long *she* would be *cross* and *flustered, cooling/herself* with a *peacock/feather/fan* until the *handle* snapped in two. The *servants* were careful to keep out of *her way*, and even the *king* was *upset. He/thought* and *thought* about what to do until *his* forehead was *striped* with wrinkles. (Critical items are italicized.)

Finally, it was frequently observed that which pictures Ss tended to look at in response to the words they heard seemingly indicated the interpretation they were giving to words which were ambiguous in their reference. For example, in connection with the phrase "birds feasting on bunches of wild grapes" (Story 3), after fixating the peacock upon hearing part or all of the word *birds*, (a) four Ss directed their fixation away from the peacock, but then back to it again upon hearing part or

all of the word *feasting*, (b) five Ss directed their fixation away from the peacock to the grapes, upon hearing part or all of the word *feasting*, and (c) other Ss indicated a failure to resolve the ambiguity of reference in the subject–object oriented term *feasting*, by rapidly vacillating their fixation between the peacock and the grapes.

DISCUSSION

The results of this study provide evidence in support of the main hypothesis and its application as a new method for testing language comprehension. Furthermore, it was seen that people sometimes obey a negative reformulation of this hypothesis by executing withdrawal responses to the initial phonemes of pronounced words. Summarized below are some of the important characteristics of this response system, certain methodological considerations to take into account when applying this response system and its characteristics as a new research tool, and specific indications of how the technique of correlating eye fixations with the meanings of concurrently heard words could potentially be applied to the detailed investigation of speech perception, memory, and language processing.

Characteristics of the Response System

Linguistic sensitivity, short latencies, spontaneity. The results indicate that Ss correctly responded not only to words representing many different grammatical functions (e.g. nouns, pronouns, adjectives, verbs, etc.), but also to the interpretations of these words as they appeared in context. Additionally, Ss frequently fixated targets prior to the complete mention of their names or other verbal referents, as anticipatory responses to cues extracted from immediately prior words (e.g. adjectival modifiers), or else from the accumulated phonemes and syllables of the current word (e.g. fixating a zebra, lion, or queen upon hearing *ze*, *li*, or *quee*, respectively). The existence of an anticipatory characteristic implies that people undergo an active online process of constructing hypotheses regarding the next successive informative item of speech, and then use the visual–motor system to test out those hypotheses prior to confirmation.

Preliminary indications reveal an automatic (not consciously planned) character to this response system. This is based on (1) its sensitive anticipatory characteristic, (2) the behavioral sensitivity displayed in the use of context cues to execute correct responses to pronouns and other contextually dependent lexical items, as well as in the formation of a large number and variety of indirect associations between spoken words and pictures, (3) the small magnitude of associated response latencies,

and (4) Ss' post-experimental verbal reports (when asked whether their correlative fixation behavior was purposeful or automatic, over 85% of Ss tested characterized this behavior as an automatic tendency triggered by concurrently heard words, at the same time strongly denying any conscious attempts to please the experimenter).

Methodological Considerations

The characteristics of this response system recommend its use as a sensitive response measure for laboratory studies in perception and cognitive psychology, with significant advantages over currently employed procedures. However, one should take into account the following considerations.

Although the percentages for experimental Ss given in Table 1 are relatively low (e.g. 37% correct fixation responses for the direct noncontextual category), this is because (1) many fixation responses which may have been functionally correct did not fit the present restrictive definition of correct responding and were, therefore, ignored in scoring, and (2) some Ss vacillated their fixation behavior between correlating words with pictures and failing to do so by passing into a point fixation or free scanning mode (see additional observations of Results section). Hence, the percentages in Table 1 do not reflect the fact that some experimental Ss appeared to sensitively visually select or withdraw from targets in response to words they heard throughout all four stories and both comprehension tests. In order to increase the reliability and efficiency with which the eye-movement response system executes correct fixation and withdrawal responses, it is suggested that those parameters which could potentially affect responding (e.g. distribution of attention between the visual and auditory modalities, period of prior visual exposure, slide and language presentation rates, number of objects pictured on each slide, type of prior instruction, etc.) have their values adjusted so as to optimize performance according to a criterion of (1) maximizing the frequency of correct fixation responses and (2) minimizing the corresponding latencies.

Potential Applications

Speech perception and memory. Yarbus (1967) has shown that Ss are unable to alter the direction of saccadic responses throughout their duration. It follows from this together with the present findings that (1) when there are no prior anticipatory cues, the moment of word or phoneme recognition (i.e., retrieval of item information from long-term memory) may be estimated by the time of initiation of the corresponding saccadic response (e.g. if there were no prior anticipatory cues, initiating fixation

of a zebra immediately upon hearing *ze* would imply that the moment of anticipatory recognition of the word *zebra* occurred at some point during the pronunciation of its first syllable), and (2) sensitive upper bounds on accessing (recognition) times (i.e., the duration of the memory search process) are obtainable by subtracting the time of initiation of the triggering word from the time of initiation of the corresponding fixation response. One of the purposes in carrying out such a study would be to estimate the minimal set of phonological cues required for phoneme and word recognition (i.e., the smallest sequence of cues sufficient to activate a search process through long-term memory for the anticipated word or phoneme).

Language processing. Currently employed procedures aimed at determining the number and kinds of interpretations given to ambiguous language have been largely confined to the case of complete printed sentences, where the interpretations selected by Ss have been communicated to the experimenter through post-facto verbal reporting and stopwatch measurement of verbal response latencies. These methods are (1) largely insensitive to the detailed perception of ambiguities in meaning from word to word and within a given word, (2) subject to short-term memory losses and, hence, possible confounding resulting from the relatively long latencies required to execute verbal responses (i.e., during the time required to initiate a verbal report Ss could forget which interpretations they had constructed as the sentence was being read, or verbalize interpretations different from those which were available as the information was being processed), and (3) additionally constrained by implicitly requiring that Ss be aware of the manner in which they respond (i.e., the number and kinds of interpretations actually available to Ss could exceed or be different from those which are theoretically susceptible to being verbally reported). Furthermore, possible influences of the contemporary visual field on the manner in which spoken language is processed have been ignored.

Because the eye-movement response system in the presence of ongoing heard language is at times characterized by a high degree of linguistic sensitivity and small latencies (including a built-in anticipatory characteristic), the present technique of correlating the visual selection of appropriate targets with concurrently heard words could potentially be applied to study in great detail the manner in which people interpret and process spoken language in the context of their contemporary visual field. If each spoken phrase was accompanied by a simultaneously presented set of pictures depicting all of the cumulative and individual word interpretations of that phrase, then the specification of the nature and order of pictures fixated, together with the associated latencies and

fixation times, would be useful in ascertaining and studying the instantaneous interpretations given to referentially ambiguous terms, as well as for studying the evolution of these interpretations as the language is being processed.

REFERENCES

CARPENTER, P. A., & JUST, M. A. Semantic control of eye movements during picture scanning in a sentence-picture verification task. *Perception & Psychophysics*, 1972, **12**, 61–64.

FENDER, D. H. Control mechanisms of the eye. *Scientific American*, 1964, **211**, 24.

MACKWORTH, N. H. The wide angle reflection eye camera for visual choice and pupil size. *Perception & Psychophysics*, 1968, **3**, 32–34.

MACKWORTH, N. H., & MORANDI, A. J. The gaze selects informative details within pictures. *Perception & Psychophysics*, 1967, **2**, 547–551.

NOTON, D., & STARK, L. Eye movements and visual perception. *Scientific American*, 1971, **224**, 35–43.

WILLIAMS, L. G. The effect of target specification on objects fixated during visual search. *Perception & Psychophysics*, 1966, **1**, 315–318.

YARBUS, A. L. *Eye movements and vision*. New York: Plenum Press, 1967.

YOUNG, L. Measuring eye movements. *American Journal of Medical Electronics*, 1963, **2**, 300–307.

(Accepted August 23, 1973)