major differences between the eye movements of addicts and controls were reflected in fixation frequency and duration, not in their task performance as measured by the subject's recognition memory for previously seen words and objects. Such results suggest that eye movements (fixation number and duration) may be a sensitive measure of gating and timing processes that group incoming sensory data, and that these data are not easily detected by conventional performance measures, such as, the percentage of targets seen and identified.

VII.4

A Framework for a Theory of Picture Recognition[1]

Geoffrey R. Loftus

University of Washington

As suggested by the title, I do not intend to present a full-blown theory of picture recognition. Rather, I want to suggest a framework within which such a theory might be couched, concentrating primarily on how information from a visual scene is encoded. Within this framework, eye fixations play a dual role. First, the pattern of eye fixations over a picture provides a powerful, apparently valid, overt measure of the parts of a picture to which the observer is attending. Second, it is suggested that the processes involving acquisition of information within a single eye fixation should be viewed as a central component in any theory of picture memory.

Picture Recognition versus Recognition of Verbal Material

Since we currently have some fairly sophisticated theories of recognition memory for verbal material (Bernbach, 1967; Kintsch, 1970; Anderson & Bower, 1972), I'd like to start with some preliminary remarks on the question of why a theory dealing specifically with picture recognition is needed in the first place. To answer this question, I'll discuss what I consider to be two fundamental differences between verbal stimuli (e.g., words, digits, letters, etc.) and pictorial stimuli. I will then argue that these *stimulus* differences have some logical implications vis-à-vis *processing* differences.

What stimulus is processed? Consider a verbal stimulus, such as a word which a subject knows he will eventually be asked to remember. Common to most

[1] This research was supported, in part, by National Science Foundation Grant GB39615 to the author.

theoretical frameworks is the notion that after a fairly early processing stage, the continued physical presence of the to-be-remembered stimulus becomes unnecessary. This is because a pattern-recognition process is assumed to operate on the physical stimulus which results in the activation of some preformed representation of the stimulus from long-term store. Subsequent processing may then be done on this representation rather than on the physical stimulus itself. A picture, on the other hand, is a genuinely "new" stimulus in the sense that the observer has presumably never seen it before. Lacking a preformed representation of the picture, all processing must be done on the physical stimulus itself. This fact has at least two implications. First, if the physical stimulus is removed, processing must halt. This notion has been confirmed in several experiments (Potter & Levy, 1969; Shaffer & Shiffrin, 1972; Loftus, 1974). The second implication is that, when presented with a new picture, the observer is faced not only with the task of encoding information about the picture into long-term memory, but also with the task of "exploring" the picture in order to decide which aspects of it deserve attention (foveal processing) and which do not. (This is not to say that a period of exploration is followed by a period of encoding. Probably both processes are carried out simultaneously.)

Time-tag information versus isomorphic-stimulus information. In theories and experiments dealing with memory for verbal stimuli, the physical form of the stimulus generally assumes little, if any, importance. That is to say, the to-be-remembered information can be presented to a subject in an infinite variety of physical forms. It can be presented visually in any number of writing styles or type faces; it can be presented auditorially by a man or a woman or a computer or a parrot. It is then typically the case that the memory test does not require information about the physical form of the stimulus, but rather only the information that the stimulus, in some physical form or another, occurred in the study phase of the experiment. Therefore, at the time the information is originally presented, it is the task of the subject simply to tag it as having occurred at that particular time (cf. Anderson & Bower, 1972). Again, the situation is quite different when pictures rather than verbal material are used as stimuli–here, the physical form of the stimulus is of paramount importance. Suppose, for example, that during a picture-recognition test, an observer is looking at a picture of a mountain, trying to decide whether to classify the picture as "old" or "new." The observer's decision is one of whether or not he has previously seen the *identical physical* stimulus which he is now looking at. This means that encoding of a picture should consist, at least in part, of formulating a memorial representation of the picture which is, in some sense, isomorphic to the original stimulus–as opposed to simply *tagging* some already-stored information as having occurred in the experimental situation.

Bearing all this in mind, let me now proceed to my vision of what a theory of picture recognition might look like. Figure 1 sketches this framework, which is initially broken down into components involving encoding processes (top box) and components involving response processes (bottom box).

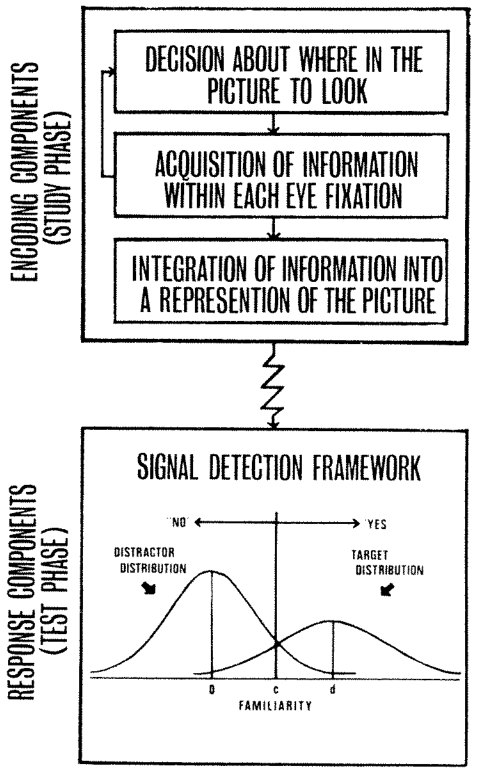

ENCODING COMPONENTS (STUDY PHASE)
DECISION ABOUT WHERE IN THE PICTURE TO LOOK
ACQUISITION OF INFORMATION WITHIN EACH EYE FIXATION
INTEGRATION OF INFORMATION INTO A REPRESENTION OF THE PICTURE
RESPONSE COMPONENTS (TEST PHASE)
SIGNAL DETECTION FRAMEWORK
"NO"
"YES"
DISTRACTOR DISTRIBUTION
TARGET DISTRIBUTION
0
c
d
FAMILIARITY

FIG. 1. A general framework for a theory of picture memory.

Response Processes

At the time a subject is deciding whether to respond "old" or "new" to a test picture, a great deal of processing is undoubtedly going on. However, due to time and knowledge limitations, I'm going to give short shrift to response processes and say only that I believe that at a rather general theoretical level, the theory of signal detection provides a good working framework. The application of signal detection theory to recognition memory is well documented (e.g., Egan, 1958; Kintsch, 1968; 1970; Freund, Loftus, & Atkinson, 1969) and I won't dwell on it here except to make a few brief comments on the construct of "familiarity." In terms of ultimately generating a more specific theory of picture recognition, it seems reasonable to postulate visual "features" which may be extracted from a picture at the time it is originally viewed. Then, instead of talking about distributions of "amount of familiarity" possessed by target and distractor pictures, we can talk about distributions of "numbers of features." I prefer to think of things this way because a feature seems to be a somewhat more tangible entity than a "unit of familiarity" as a candidate for something that can be extracted from a picture. If we talk about features from pictures, then we can also talk about (1) distributions of number of features that may be extracted during an eye fixation and (2) sets of features that are shared by targets and distractors. This second notion could serve to clarify the effects on recognition performance of target–distractor similarity.

Encoding Processes

Three encoding processes have been included in the top box of Fig. 1. First, as noted earlier, a decision must be made as to which parts of the picture should be processed (attended to). At any given time, this corresponds to a decision about where the next eye fixation should be. Second, once a particular area of the picture is being fixated, information must be extracted from that area and processed during the fixation. Finally, the information extracted during a series of fixations must be integrated into some over-all representation of the picture.

Where to look. Since the pioneering work of Buswell (1935) it has been clear that eye fixations are not distributed randomly over the picture. Rather, a large majority of the fixations are made on a rather small number of "areas of general interest" in the picture. This makes intuitive sense. If I show you a picture of the New York skyline under a clear blue sky, you are more likely to fixate on, say, the Empire State Building than somewhere in the middle of the clear blue sky.

Following Buswell's work, there have been attempts to specify the notion of an "area of general interest" somewhat more precisely. Berlyne (1958) presented subjects with pairs of pictures of the sort shown in Fig. 2. In each case, one member of the pair was defined as "informative" (in an information–theoretic sense) whereas the other member of the pair was defined as relatively less informative. Subjects tended to spend more time looking at the informative as opposed to the noninformative member of the pair. More recently, the work of Mackworth (Mackworth & Morandi, 1967; Mackworth & Bruner, 1970; cf. also Pollack & Spence, 1968) has dealt with the notion of informative areas within a complex, naturalistic scene. The procedure used in these experiments was to divide a picture into an 8 X 8-in grid. A group of subjects then rated how informative was each individual square, following which an independent group of subjects was permitted to view the entire picture. Eye-fixation patterns were recorded from the second group, and the results indicated a strong positive correlation between the informativeness rating of a particular square and the number of fixations made on that square.

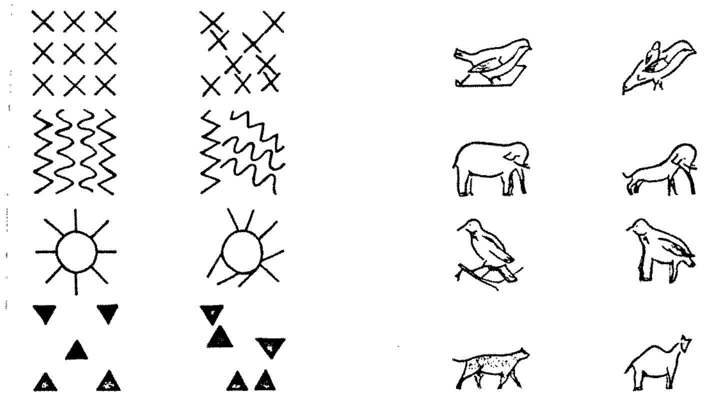


FIG. 2. Stimuli used in the Berlyne (1958) experiments. For each pair, one member is defined as "informative" (in an information–theoretic sense) whereas the other member is noninformative.

Mackworth's work has been valuable in confirming the existence of informative areas in pictures, but his definition of *informativeness* is highly empirical—an area is informative to the extent that other people say it is informative. Combining Mackworth's results with those of Berlyne, however, it seems reasonable to expect that subjects tend to look at areas of pictures which may be specified a priori as being informative in an information–theoretic sense. More precisely, I would like to offer the following definition of an informative area (or object) in a picture: An object in a picture is informative to the extent that it has a low conditional probability of being there given the rest of the picture and the subject's past history. As an example of what I mean by this, consider Figs. 3 and 4. Figure 3 shows a picture of a farm and contains a number of objects—the farmhouse, wagon, tractor, etc.—which most of us would agree belong on a farm. Figure 4 depicts exactly the same scene with one exception: an octopus has been substituted for the tractor. According to my proposed definition of informativeness, this octopus would constitute an informative object to any person whose experience with farms has not included the presence of an octopus.

If subjects do, in fact tend to fixate on areas of pictures which are informative by this definition, then they are carrying out the most efficient strategy possible in terms of subsequently being able to recognize the picture. Recognition of a picture involves being able to *discriminate* the picture from other similar pictures.[2] Therefore, the most valuable aspects of a picture to encode are those

[2] This notion naturally assumes that the picture being viewed is a member of some known, reasonably well-defined class of pictures. In a typical picture-recognition experiment, the class quickly becomes apparent to a subject via experimental instructions, warm-up pictures, or the first few pictures of the study sequence. Thus, the class of pictures might be naturalistic scenes, faces, common objects, etc.

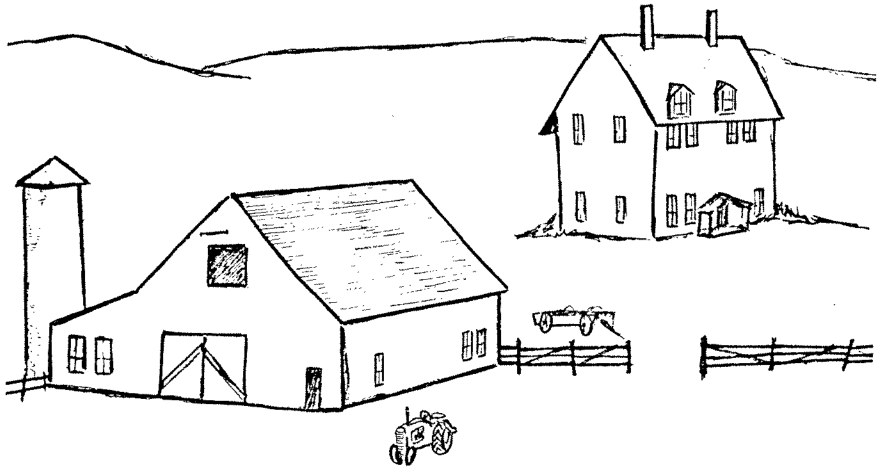


FIG. 3. A picture containing several noninformative objects.

aspects that are least likely to be common to other pictures being viewed. If a subject were looking at Fig. 4 trying to encode it for subsequent recognition, he would be in good shape by encoding the presence of an octopus, since any potential distractor pictures of farms would be unlikely to contain an octopus—i.e., the octopus provides the best discriminative cue.

These speculations suggest an obvious experiment that we are currently carrying out. We have created a large number of pairs of picture similar to the pair of Figs. 3 and 4. Each subject views a series of pictures, half of which contain an informative object (like Fig. 4) and the other half of which do not contain an informative object (like Fig. 3). Eye fixations are recorded during initial viewing and the pictures are later tested in a yes–no recognition procedure. The data have not yet been collected, so I have created them instead. Figure 5 shows the pattern of results we expect (i.e., hope) to get. The top panel shows the cumulative probability of having fixated an object as a function of the ordinal fixation number on the picture. If my definition of an informative object corresponds to what Mackworth's subjects called an informative object, then a given object in the picture should be fixated sooner when it is informative than when it is not informative.

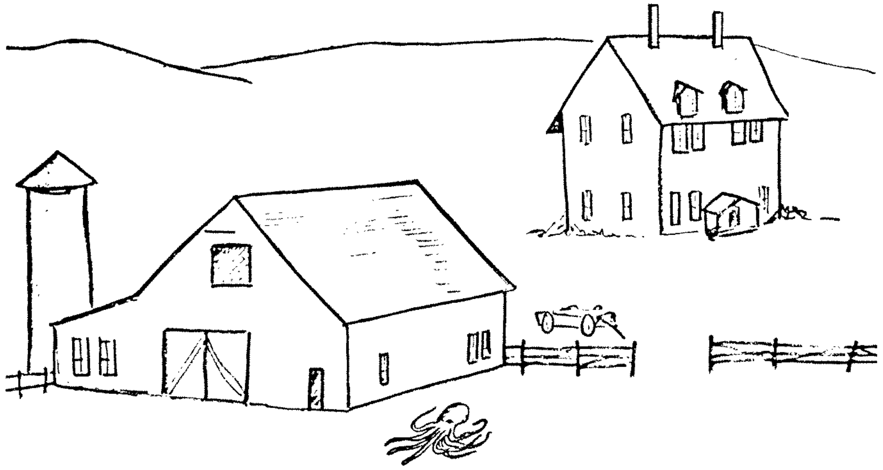


FIG. 4. A picture which contains one informative object—the octopus.

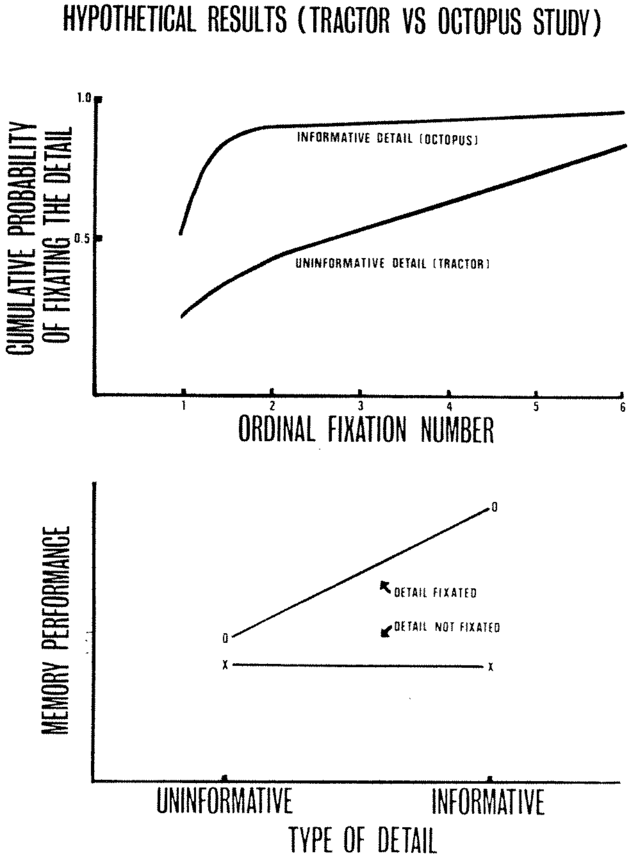
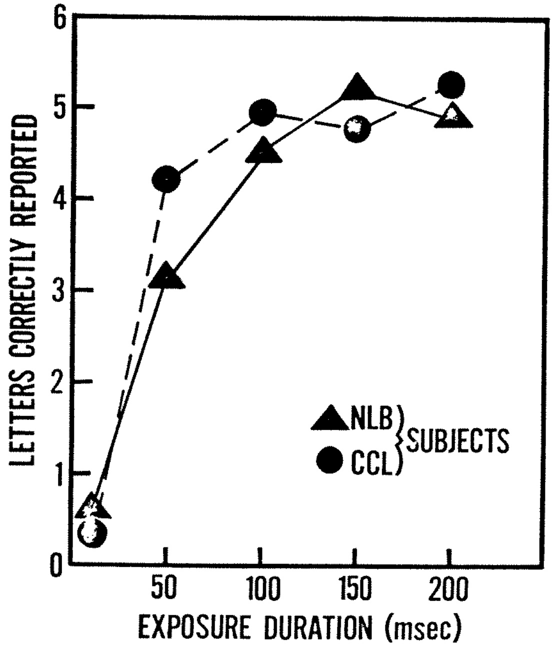


FIG. 5. Hypothetical results. The top panel depicts expected results for eye movements and the bottom panel depicts expected results for recognition performance.

The predicted relationship between informativeness and subsequent recognition performance is depicted in the bottom panel of Fig. 5. When an informative detail is fixated, it should aid recognition performance relative to cases where an informative detail is not fixated or where the fixated detail is not informative.

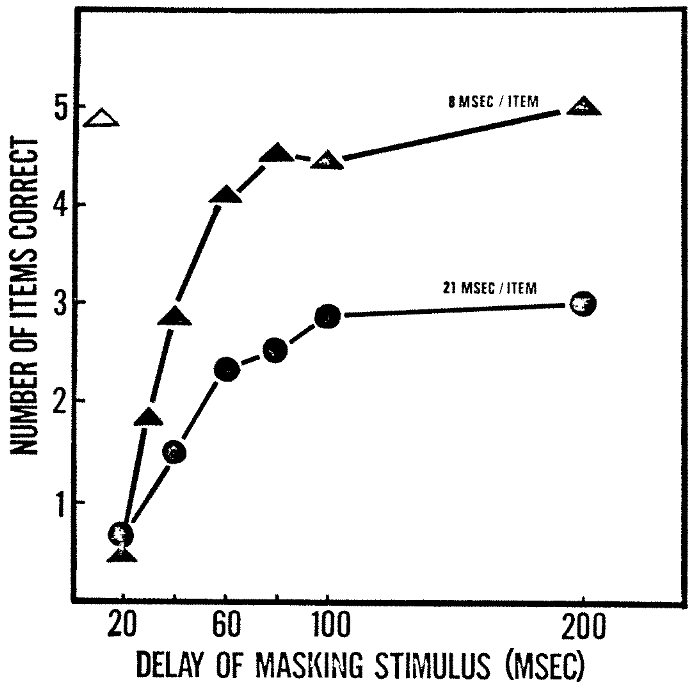
Processes occurring within a fixation. We now arrive at the second encoding component. Having decided where to fixate and having fixated there, the observer must now extract information from the fixated area. A good place to go for clues as to what is happening within a fixation is the voluminous literature on the information available within a single brief visual presentation. Research in this area has proceeded under the rationale that an understanding of the processes taking place during a controlled tachistoscopic presentation will in turn provide an understanding of the processes occurring within an eye fixation. Indeed, a classic paper by Sperling (1960) begins with the statement, "... [the question of how much can be seen in a single brief exposure] is an important problem because our normal mode of seeing greatly resembles a series of brief exposures ... [and] the eye assimilates information only in the brief pauses between saccadic movements [p. 1]."

A paradigm that simulates a single eye fixation using a tachistoscope is one introduced by Sperling (1963). This paradigm involves the presentation of an array of verbal stimuli (e.g., letters) for a brief, variable amount of time, followed by a visual noise mask. Figure 6 shows Sperling's results. The amount of information acquired from the array (as measured by the number of letters reported) increases with the exposure time of the array up to about 100 msec and then asymptotes. A question of some potential importance is: Why does there appear to be no further acquisition of information after 100 msec? One possibility is that the onset of a new visual stimulus activates a pattern recognizer—or more generally, a visual information-acquisition process—that operates for only about 100 msec following the onset of a visual stimulus and then stops and is idle until the eye is presented with a new stimulus. Carrying this notion over to an eye fixation, this would mean that within a given eye fixation, information from the stimulus being fixated would be acquired for only the first 100 msec or so following the onset of the fixation.



(AFTER SPERLING, 1962)

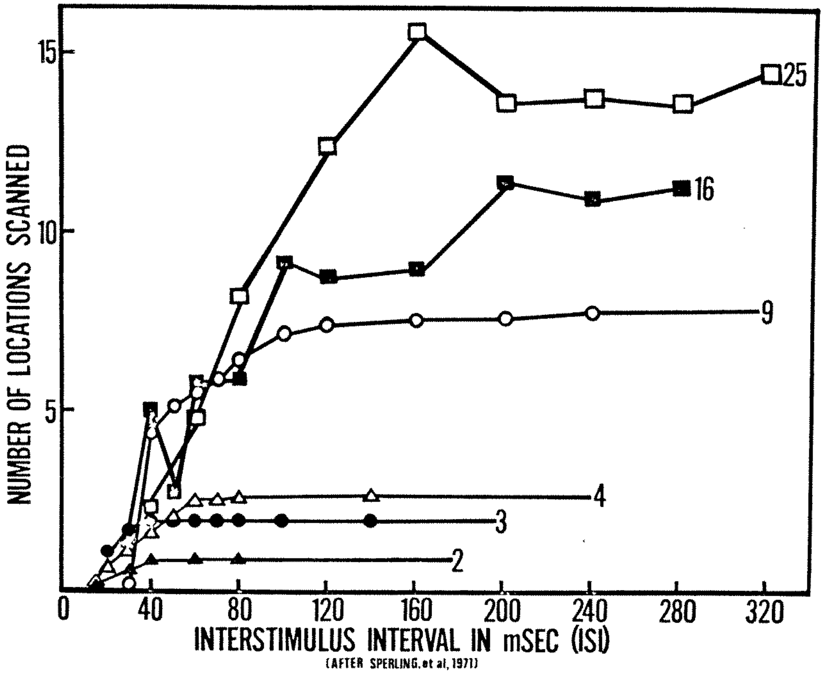
FIG. 6. Memory performance (number of items correctly reported) as a function of exposure time of the letter array. (After Sperling, 1963.)



(AFTER ALLPORT, 1968)

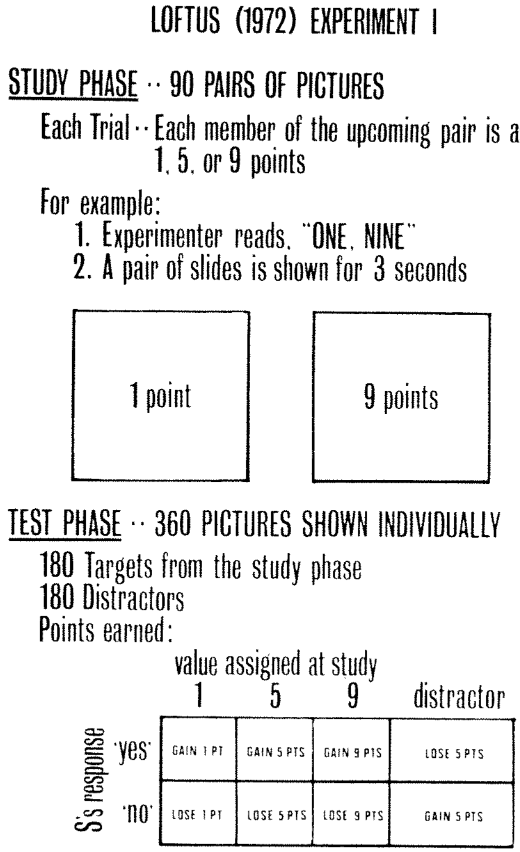
FIG. 7. Memory performance (number of items correctly reported) as a function of the exposure time of the stimulus array. The top curve represents data when letters are used as stimuli and the bottom curve represents data when Landolt C's are used as stimuli. (After Allport, 1968.)

A second somewhat less interesting explanation for the asymptote in Fig. 6 is that the five items acquired in the first 100 msec fill up short-term store. However, other data do not support this possibility. Figure 7 shows data collected by Allport (1968). Allport used the same paradigm as did Sperling but used two types of stimuli. The top curve in Fig. 7 shows the results when letters were used as stimuli whereas the bottom curve shows the results when Landolt C's were used. If the asymptote were due to a filling up of short-term store, then it is difficult to imagine why different stimuli would produce different asymptotic levels, in view of the fact that the number of items which can be held in short-term store is relatively independent of what the items are (Miller, 1956). Even more compelling data have been gathered by Sperling, Budiansky, Spivak, and Johnson (1971). In their experiment (which simulates a series of eye fixations) subjects were shown a series of letter arrays which appeared in rapid succession on a cathode-ray tube. The size of the arrays varied from 2 to 25 letters, and the time each array remained on the screen varied from 10 to 320 msec. Embedded somewhere in one of the arrays was a digit, and it was the subject's task to report the digit's location. Using this procedure, it is possible to estimate the number of locations scanned in each array. Figure 8 shows this measure of visual information processing as a function of how long each array was presented (labeled ISI). Again, these functions all asymptote at around 100 msec. Since this paradigm almost completely eliminates short-term memory limitations, these results support the notion that the asymptote is due to a limit in how long the "visual information acquisition" program will operate following the onset of a new visual stimulus.



(AFTER SPERLING, et al, 1971)

FIG. 8. Estimated number of locations scanned as a function of exposure time of the stimulus array (ISI). The curve parameter is the number of letters in each array. (After Sperling et al., 1971.)



LOFTUS (1972) EXPERIMENT I

STUDY PHASE ·· 90 PAIRS OF PICTURES

Each Trial ·· Each member of the upcoming pair is assigned
1, 5, or 9 points

For example:
1. Experimenter reads, "ONE, NINE"
2. A pair of slides is shown for 3 seconds

1 point

9 points

TEST PHASE ·· 360 PICTURES SHOWN INDIVIDUALLY

180 Targets from the study phase
180 Distractors
Points earned:

S's response	1	5	9	distractor
	value assigned at study			
'yes'	GAIN 1 PT	GAIN 5 PTS	GAIN 9 PTS	LOSE 5 PTS
'no'	LOSE 1 PT	LOSE 5 PTS	LOSE 9 PTS	GAIN 5 PTS

FIG. 9. Design of Experiment 1 of the Loftus (1972) study.

A series of picture-recognition experiments that I have reported (Loftus, 1972) provides evidence that the information acquisition process within an eye fixation follows the same time course as that depicted in Figs. 6–8. Figure 9 shows the design of Experiment 1 of this study which was originally motivated by the question: What is the relationship between the number of fixations made on a picture and subsequent recognition-memory performance for that picture? In an initial study phase of the experiment, subjects were shown 90 pairs of pictures for 3 sec per pair. Eye fixations were recorded during this study phase. To gain some control over the number of fixations per picture, each member of the pair was assigned a value of 1, 5, or 9 points prior to the onset of the picture. This value was directly related to the amount of money the subject would gain if, during the subsequent recognition test, the subject correctly recognized the picture. It was thus expected that more fixations would be made on high-valued pictures than on low-valued pictures.

The results of this experiment showed that both number of fixations and subsequent recognition-memory performance were increasing functions of value. Figure 10 shows memory performance (hit rate) as a function of the number of fixations made on the picture at the time of study. The curve parameter is the value of the picture. Two aspects of these results are of interest. First, the more fixations accorded the picture, the higher is subsequent recognition performance. Second, with the number of fixations held constant, memory performance is independent of the picture's value. The implications of this result is that the higher memory performance on higher-valued pictures is completely mediated by the greater number of eye fixations on these pictures.

Table 1 shows average fixation duration as a function of the number of fixations on the picture and of the picture's value. Of interest is the fact that (for unknown reasons) the greater the value of the picture, the longer was the average duration of fixations made on the picture. However, as Fig. 10 shows, the extra time per fixation on the high-valued pictures did not add anything in terms of memory performance. Making the reasonable assumption that memory performance reflects the amount of information extracted from the picture, we are left with the conclusion that no extra information was acquired in the additional time per fixation on the higher-valued pictures. This result suggests that a hypothetical function relating the amount of information acquired to fixation duration would resemble the curve shown in Fig. 11. The correspondence between Fig. 11 and Figs. 6–8 should be fairly obvious—it appears to be the case that the same information-processing mechanisms operate following the

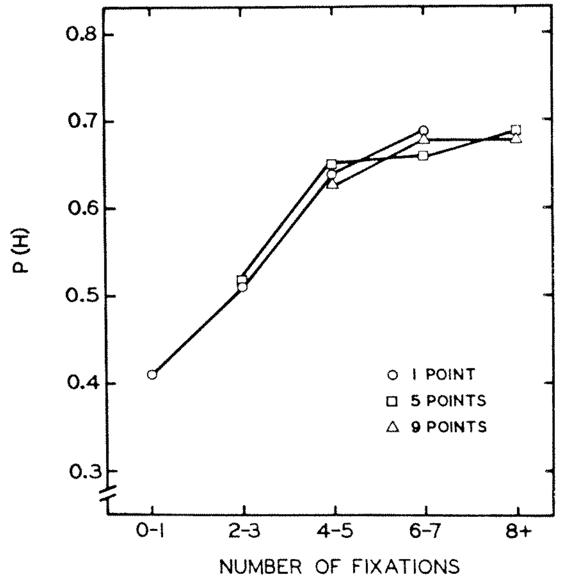


FIG. 10. Memory performance (hit rate) as a function of number of fixations accorded the picture at time of study. The curve parameter is the value of the picture.

TABLE 1
Average Fixation Duration as a Function of Value for Pictures over *i* Fixations

	Average fixation duration (sec)				
Value (points)	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$
1	.292 (117)[a]	.292 (111)	.290 (69)	.279 (49)	.300 (26)
5	.325 (85)	.311 (92)	.312 (93)	.304 (88)	.300 (69)
9	.350 (35)	.369 (59)	.336 (92)	.308 (111)	.311 (25)

[a] Numbers in parentheses are the sample sizes for each cell.

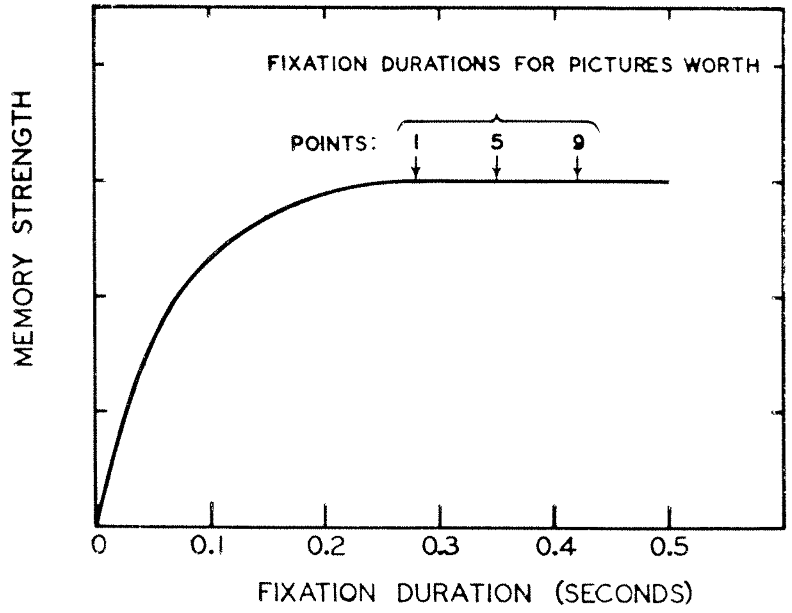


FIG. 11. A hypothetical function representing acquisition of information as a function of fixation duration for a single eye fixation.

onset of a new visual stimulus either when the visual stimulus is initiated by the observer (with an eye movement) or when it is initiated by the experimenter (with a tachistoscope).

Experiment 2 in the Loftus (1972) study provides some confirmation of this notion. In Experiment 2, single pictures were displayed in the study phase at exposure times varying from 300 to 5000 msec. Again, eye movements were recorded during the study phase, and the pictures were subsequently tested in a yes–no recognition test. Two results of interest emerged from this study. First, with exposure time held constant, memory performance was a strongly increasing function of number of fixations made on the picture at study–for example, after a 3-sec exposure, fifteen 200 msec fixations produced considerably better performance than ten 300-msec exposures. The second result was that with the number of fixations held constant, performance was independent of exposure time; for example, if 12 fixations were made during a 3-sec exposure, performance did not differ from the case when 12 fixations were made during a 5-sec exposure. Taken together, these results suggest that each fixation results in the acquisition of one "chunk" of information about the picture–and within a fixation, all the information germane to subsequent recognition memory is acquired rather quickly following the onset of the fixation.

The most intriguing question to come out of all this is: Why does it seem that the last part of each fixation is wasted time? One possibility which has been suggested by Gould (1969) is that the last part of a given eye fixation is spent computing where the next fixation will be made. I believe that this is an appealing possibility. Under this view, an eye fixation would be divided into two major (possibly overlapping) stages. The first stage would involve wide, peripheral processing to determine where the next fixation should occur. The second stage would probably be somewhat task specific (cf. Yarbus, 1967). Thus, for example, in scanning a picture, informative areas would be identified; in a visual search task, a potential target would be sought out, etc.

Integration of Information over Successive Fixations. Information integration seems to be a two-stage process. The first stage is *getting the big picture*. At least three lines of research have suggested that within a very short time after a picture has appeared, an observer has some notion of what the "gist" of the picture is. First, Mackworth's work, already discussed, shows that observers look at "informative areas" very quickly–within the first one or two fixations on the picture. In order to do this, some processing must have taken place to provide information about which areas of the picture are informative to begin with. Second, Potter (1972; has reported an experiment in which a series of pictures is shown in rapid succession (e.g., one picture per 100 msec). Subjects were instructed to press a button when they saw a picture whose gist was defined in some very vague way (for example, "a picture depicting a game") and were able to do this with no difficulty. Finally, an experiment by Biederman (1972) utilized a procedure in which subjects were shown a picture for a brief period of time. Following the exposure, they were asked to name an object from the picture whose location was specified by a visual marker. There were two conditions: in a "jumbled" condition, the picture was cut into six sections which were spatially scrambled, thereby making identification of the gist very difficult. The second condition was a control condition in which the normal, complete, unscrambled picture was shown. Object detection in the jumbled condition was poorer than in the normal condition, suggesting that rapidly acquired information about gist was being used to aid detection.

The second stage of information integration is *getting informative details*. It seems likely that, following the early acquisition of this "gist information," successive eye fixations are utilized to acquire information from informative areas in the picture, as discussed above. To make a stab at exploring this process, Susan Bell and I have conducted a picture-recognition experiment in which subjects were asked, at the time of recognition, to identify the bases of their responses. Specifically, they were asked to make one of two choices: (1) they were responding because they remembered some specific detail in the picture or (2) they were responding merely on the basis of the "general familiarity" of the picture. Exposure time was varied from 60 to 500 msec at the time the pictures were originally viewed. Several results of interest emerged from this experiment. First, consider the function relating the probability of naming a specific detail from a picture to the original exposure time of the picture. Ninety-six percent of the variance in this function was accounted for by a model which assumed that, during each eye fixation, the probability of an informative detail is encoded with some constant probability α. Second, when a detail was named, performance (measured in terms of d') was increased by about 1.5 relative to when a detail was not named.

To recapitulate, it appears that some general information about a picture is acquired very quickly after the picture is first exposed. Following the acquisition of this initial information, the task of each eye fixation is to encode more precise information about what is in the picture. In terms of picture *recognition*, a simplistic view–but one that seems to work–is that with each eye fixation, there is some constant probability that a detail will be encoded which will serve to distinguish the picture from other pictures and upon which a recognition response may be based.