THE PUPILLARY RESPONSE TO RETRIEVAL OF INFORMATION FROM LONG-TERM MEMORY

By

DONALD BRUCE HEADLEY

Bachelor of Science
University of Massachusetts
Amherst, Massachusetts
1968

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1971

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
July, 1973

THE PUPILLARY RESPONSE TO RETRIEVAL OF

INFORMATION FROM LONG-TERM MEMORY

Thesis Approved:

Robert F Stanners

Thesis Adviser

Larry T. Brown

Robert J Weber

Lyle D. Broemeling for D. L. Weeks

D D Durham

Dean of the Graduate College

ACKNOWLEDGMENTS

This research project was greatly aided by the guidance and support of my thesis adviser and committee chairman, Dr. Robert F. Stanners, who also provided the research facilities. The advice given by the other members of my committee is also appreciated: Dr. Larry Brown; Dr. Robert Weber; and Dr. David Weeks, who gave valuable statistical consultation.

Acknowledgment is also extended to Dr. William Clark and Dr. David Johnson, who had previously constructed the pupillary apparatus, and to Dr. Clark and Dr. Stanners for development of a computer program for pupillary data.

This project was supported in part by funds from a National Science Foundation Traineeship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

LITERATURE REVIEW

Memory involves the acquisition, storage, and retrieval of information. This paper presents an experiment which was concerned with the retrieval of knowledge from long-term memory (LTM). The basic issue was to study how people recall facts which they have previously learned. A review of the theoretical issues will be presented first, followed by a discussion of pupillometry, a technique which provided an index of cognitive processing as subjects attempted to answer questions dealing with general knowledge.

Theories of Retrieval

Shiffrin and Atkinson (1969) and Shiffrin (1970) propose that retrieval involves three primary mechanisms. To begin with, they envision LTM to be a "self-addressing" system: a question put to the system places constraints on what memory areas will contain the desired information; a search process examines the most likely locations (natural starting points), and produces images ("inter-associated complexes of information"). A recovery process may then transfer some of this information into a short-term store, or "working memory". Finally, a response generation component of retrieval makes one of three decisions: emit a response, abandon search, or continue the search process (time limitations may of course limit the extent of the recursiveness of the search component). If search is continued, a new location may be examined, and if an image is produced, the information recovered from it will be weighed for possible output. Retrieval thus involves not only recovery of information, but also decisions concerning response production.

Norman (1970, 1972; Norman & Rumelhart, 1970) postulates that the act of remembering consists of three stages. The preprocessing stage tests the legality of the input; if processing of a "query" puts it "into the proper format for entrance into his memory schema", (Norman, 1970, p. 30) then the second stage, memory search, occurs. This process makes use of immediately accessible "landmarks" of memory, and retrieves all information relevant to the reformulated version of the query. The post-processing stage acts upon the retrieved information; this last stage is a decisional process and includes the biases and criteria of the subject (S) and judges the accuracy of one's potential response.

Norman feels that the storage system is content-addressable (a less specific form than self-addressable): a query initiates a rapid search to determine whether any information about it exists in memory (preprocessing); if so, memory search utilizes the contextual information in the question to gain access to the relevant attributes associated with this particular context. Thus, search recovers those locations in memory containing the context of the question. Then decisions are made on the retrieved attributes.

This judging of potential responses is called the "editing" process and is a feature of most theoretical accounts of recall. For example, Kintsch (1970, pp. 273-277) proposes that an implicitly retrieved response is checked (recognition process) for such criteria as familiarity or context before it is either rejected or made overt. A recall model by

Anderson and Bower (1972) has similar notions. They argue that recall attempts initiate a directed search process. A retrieval mechanism generates implicit "candidates" for possible recall. These covert responses are input to a recognition process which edits the information; a decision at this stage is based upon the retrieval of contextual information (a unique feature of the model is the assumption that "the processes underlying recognition in free recall are identical to those processes which underlie performance in a pure recognition task": pp. 116-117). If the context-retrieved information exceeds a criterion, the item is designated for overt recall.

Adams (1967, pp. 279-305; 1968) states that recall involves response-monitoring behavior. The process involves the subjective monitoring of implicit or overt responses, and it is responsible for the acceptance or rejection of the retrieved response. The concept is necessary, Adams postulates, to account for data which can not be handled by the conventional S-R verbal learning model (i.e., a memory trace bonds the stimulus and response, and the strength of the trace is of prime importance). An example is omission behavior -- sometimes weak habit strength is not the only reason why Ss fail to give a response for a particular item. On occasion, an implicit response being judged for recall may be subjectively rated for its correctness. At other times, an overt response may be quickly taken back, and the correct answer then readily supplied. Such response-monitoring behavior is not adequately predicted by the traditional S-R model (an "open loop" model: response either occurs or does not occur). Response recognition operates under a "closed loop" principle -- a response is compared via feedback to a reference level, that of a previously established "response-produced perceptual trace".

Tulving and Madigan (1970), in a review of memory and verbal learning, also remarked that traditional models can not adequately account for monitoring behavior:

> No extant conceptualization, be it based on stimulus-response associations or an information processing paradignm, makes provisions for the fact that the human memory system cannot only produce a leanred response to an appropriate stimulus or retrieve a stored image, but it can also rather accurately estimate the likelihood of its success in doing it... (p. 477).

One feature which characterizes human memory as an efficient retrieval system is that one can determine very quickly whether or not he has knowledge of a given event. We can rapidly reject a question as being one of which we have no knowledge. To other questions, memory locations are rapidly searched for the answer. However, successful response production may not always be the end result of a search. A person may emphatically state that he knows the answer, but for the moment it is blocked out of his memory. The search failed to locate a strong memory trace. In this case, recall failure is due to the inaccessibility of otherwise available information (this availability-accessibility distinction was made by Mandler, 1967, and has been empirically validated by Tulving & Pearlstone, 1966). Several investigators have concerned themselves with determining how accurately one can evaluate this availability of information in memory. That is, can one accurately judge the contents of his memory and predict for a given item whether he has enough information stored either to recall it, if given retrieval cues, or recognize it if he were to see or hear the correct item. The discussion will now turn to a review of this literature.

"Feeling of Knowing" Studies

The phenomenon under review may be described as follows: a person is unable to recall particular information which he once knew, but feels that he is on the verge of retrieving it, or could recall it if given time or clues. This state has been discussed under a variety of names -- sense of familiarity (James, 1899), condition of readiness (Woodworth & Schlosberg, 1954), motivated forgetting (Freud, 1955), the tip of the tongue phenomenon (Brown & McNeill, 1966; Freedman & Landauer, 1966), feeling of knowing (Hart, 1965, a, b; 1966; 1967 a, b), judgements-of-knowledge (DaPolito, Guttenplan, & Steinitz, 1968), and recall readiness (Flavell, Friedrichs, & Hoyt, 1970).

The phenomenon was discussed by William James (1899):

> Suppose we try to recall a forgotten name. The state of our consciousness is peculiar. There is a gap therein; but no mere gap. It is a gap that is intensely active. A sort of wraith of the name is in it, beckoning us in a given direction, making us at moments tingle with the sense of our closeness, and then letting us sink back without the longed-for term...The rhythm of a lost word may be there without a sound to clothe it; or the evanescent sense of something which is the initial vowel or consonant may mock us fitfully, without growing more distinct (pp. 251-252).

Woodworth and Schlosberg (1954) described this state as a "peculiar feeling of nearness"; a person feels that "... a little extra push in the right direction will give complete recall" (p. 720). Freud (1955) also discussed this temporary inability to retrieve a previously known name, and cited cases in which evoked responses, though incorrect, bore associations to the correct word in the form of sound patterns, number of syllables, and certain letters.

Although this situation of experiencing temporarily blocked information had been colorfully described, very little experimental investigation was reported until a study by Brown and McNeill appeared in 1966. They defined this "tip of the tongue" (TOT) state as one in which "recall is felt to be imminent". A subject in such a state "would appear to be in mind torment, something like the brink of a sneeze, and if he found the word his relief was considerable" (p. 326). In this study, TOT states were induced by the reading of dictionary definitions of low frequency words (e.g., nepotism, sampan). If an S could not actually supply the word, but felt that he knew it and was on the verge of retrieving it, he was to supply certain information concerning the target item. This information included the initial letter, the number of syllables, similar sounding (SS) words, or similar meaning (SM) words. The guessed number of syllables accurately corresonded with the actual number of syllables for one, two, or three syllable words. The initial letter was correct for 57% of the responses. For both the syllabic and initial letter data, however, that of the SS words was superior to that of the SM words given by the subjects. It appears that sounds of words act as more powerful mediators in word production. There was suggestive evidence from a further analysis of the SS words that the location of the primary stress was successfully identified. Finally, the letters of SS words were analyzed as a function fo serial position, and results showed that the highest percentages of matches occurred at the beginning and end of a word.

The data collected on unrecalled words was classified into two groups by the authors: abstract form recall (number of syllables, and primary stress location), and partial recall (reproduction of correct letters in various parts of the word). Both types were considered to be generic recall, "since the class of words defined by the possession of any part of the target word will include words other than the target"

(p. 326). Generic recall indicates that particular information about a word is accessible even though the specific target item itself is not. Other studies have shown that recall cues in the form of words from various positions of the target word can produce correct recall. For example, Horowitz, White and Atwood (1968) had their Ss study a word list, and then gave them letter combinations from various parts of the word as a recall-aid; the most effective word elicitors were cues from the beginning of a word, followed by those from the end and middle. Thus, the same kind of partial recall which Ss produce to TOT words serves as efficient recall cues. These data suggest that memorial representation of items is in the form of discrete features, or what Underwood (1969, p. 566) calls "orthographic attributes". A TOT state should therefore reflect the fact that a S has access to a certain number of attributes about the item.

A TOT state was indicated by Ss on 360 occasions out of 2744 possible (56 Ss X 49 words in list); 127 of these instances proved to be false alarms (Ss did not really "know" the word). There is some question, then, concerning the accuracy of TOT states. How accurately is a person able to judge the content (availability) of his memory? A study of this question must allow a sensitive means for a S to demonstrate any information he may have about a particular item. In a series of investigations, Hart (1965a, b; 1966; 1967a, b) studied the accuracy of availability-judgments by making use of the fact that items which can not be recalled can often be recognized. He termed the feeling state the "feeling-of-knowing experience" (FOK) to take into account not only the more intense TOT states, but also the "more general and ubiquitous" instances of lesser intensities, "occurring every day with many types of memory materials: names, dates, telephone numbers, addresses, faces, places, etc." (1965b, p. 208).

In the 1965b study, Experiment I investigated the accuracy of dichotomous FOK ratings ("Yes": S feels he knows enough about the answer to recognize it; "No": S predicts he probably will not recognize it). Experiment II studied graded strengths of knowing (judgment scale ranged from "definitely yes" to "definitely no"). Hart's technique was as follows -- a general information questionnaire was given to a S, and for those items not answered, he rated whether he would be able to choose the correct answer on a four-choice recognition test. The data was divided into one of four categories for analysis: an item given a FOK rating of "Yes" may or may not be correctly chosen on the recognition test; similarly, one rated "No" may or not be correctly recognized. The critical comparison was between the proportion of correct recognitions for "Yes" and "No" ratings. If the FOK experience is accurate, Ss should recognize those items they feel they know, but miss the ones they feel they do not know. The results of Experiment I confirmed this prediction; the proportion of successful recognition for Yes-rated items was .76 while for No-rated items it was .43, a significant difference.

Experiment II found a similar significant difference. The questionnaire was made more difficult by the addition of 25 questions which contained equally likely alternatives on the recognition test. It was felt that the higher-than-chance recognition of No-rated items in Experiment I was in part due to the diversity of alternatives. Results were as follows: the proportion of Yes-rated items correctly recognized was .66; the proportion of Yes-rated items not recognized was .34. This result implies that the FOK experience is an accurate indicator of what is in memory (this comparison was significant in Experiment I as well). The proportion of No-rated items not recognized was .62, whereas that for No-rated items recognized was .38. This result indicates that the rating process is also an accurate indicator of what is not in memory.

The figures given above for Yes/No judgments in Experiment II represent values pooled over the rating scale categories. Ss were asked to rate their FOK on a 3-point scale within the Yes or No category. When correct recognitions were evaluated as a function of scaled ratings, the six proportions fell into a suggestive, but non-significant, four category classification: "definitely yes", "maybe yes", "maybe no", "definitely no". The implication was that the accuracy of FOK experiences may extend over more than just a dichotomous Yes/No scale.

Hart termed the intervening process which produces FOK judgments the "memory-monitoring" (MEMO) process. The above experiments illustrate the accuracy of the process. A further study (1967b) tested monitoring accuracy for second-try recall. Ss first were given a fast-paced recall test, and for those items not recalled, were to give FOK judgments which predicted whether they could recall the item if given more time. On this second recall test, if an item still could not be recalled, Ss gave a FOK rating for predicted recognition ability; finally, the recognition test (forced 4-choice) was administered. The paradigm, then, was: Recall-FOK-Recall-FOK-Recognition. Of those items given a Yes-FOK rating after the first recall test, .25 were recalled correctly on the second test (only.05 of those rated No were achieved correctly on this test). Of those items rated Yes after the second unsuccessful recall attempt, .59 were recognized correctly, whereas .30 No-rated items were recognized correctly. This result is comparable to that in the 1965b experiment.

Hart concluded that "The monitor appears, from these results, to be a more accurate predictor of recognition performances than recall performances" (p. 196). Apparently, Ss were over-estimating their second-try recall abilities. It appears that accessibility is better when overt retrieval cues such as those on a recognition test are supplied, although implicitly generated cues do result in some successes, as evidenced by the .25 correct recall performance on the second try (In the Brown & McNeill, 1966, study, 36 of the 233 TOT instances resulted in a correct recall of the target word.).

Finally, Hart (1967a, Experiment I) tested whether FOK accuracy applies not only to knowledge concerning general information questions, but also to items which are learned in the laboratory, such as paired-associates. That is, "Will the MEMO process operate accurately for recently learned, comparatively meaningless memory items, as it does for factual questions?" (p. 686). After studying a list of 48 word-CCC (consonant trigram) pairs, Ss attempted to recall the proper trigram for each word. Unsuccessful recall was to be supplemented with a FOK judgment concerning predicted recognition success. Results were as before: Yes-FOK, judgments resulted in significantly more correct recognitions than did No-FOK judgments.

Experiment II was essentially a replication of I. As in Experiment II of the 1965b study, a 6-point FOK rating scale was used, 3 points for gradations of Yes feelings, and 3 points for No feelings. However, within either the Yes or No classification, the proportions of correct recognitions were highly similar to each other. When pooled over the 3 ratings, Yes-recognitions were again significantly greater than No-recognitions.

A procedural change was initiated in the 1967a,b studies. In a methodological note, Hart (1966, Experiment II) tested a possible bias which could have conceivably caused an inflation of the Yes-recognition proportion. If Ss were cautious and withheld answers on the recall test, even though they had a very strong suspicion of what the correct item was, many questions would have been given a Yes FOK, some of which would have been easily answered on the recognition test. Thus, when the Yes-recognized proportion was compared with the No-recognition proportion, a false indication of MEMO accuracy might be obtained. Consequently, Hart (1966) required Ss to supply a response to every question on the initial test, even if it was a guess or an association. Ss were instructed to give a FOK judgment to those items they felt were answered incorrectly. Hence, FOK-rated items should now reflect only those which were truly blocked. Results indicated that while the proportion of correctly recognized Yes items did drop somewhat, the difference between this proportion and that of correctly-recognized No items was nevertheless a significant one.

The general conclusion from the FOK articles cited above is that people do have some idea of what is in their memory even when an overt response can not be produced. The monitoring process leads to fairly successful predictions of not only what is in memory, but also what is not in memory. This process is not perfectly accurate, because if it were the Yes-recognized proportion should be 1.00 whereas the No-recognized proportion should be .25 (chance performance on a 4-choice recognition test). Hart (1965a,b; 1967a,b) has offered some speculations on the functional significance of a MEMO process. Unlike a computer, which has perfect storage and retrieval systems, the human memory has its limits. Not all information can be stored, nor can that which is stored be always retrieved. To be useful, FOK experiences must be accurate. Thus, Yes-FOKs should tell one that continued searching for particular information may produce it, and these experiences should generally culminate in a successful retrieval; a No-FOK should not occur if the searched for information could have in fact been retrieved with further persistence. Hart's studies show that MEMO is for the most part an accurate process.

Flavell et al. (1970) present data which suggest that the memory monitoring process is one which is present in children as well. They studied the (1) predicted object memory span and (2) "recall readiness" of children from four school levels: nursery school, kindergarten, second, and fourth grade. In the first task, the children predicted how many different pictures they could memorize and report in serial order. An actual memory span was then determined. In the second part, the Ss studied a number of pictures which equalled their actual object span, and indicated to E when they thought they could give perfect recall. The memory span experiment tested Ss' knowledge of their memory abilities for a future task, whereas the second study determined their ability to monitor recall readiness while they were engaged in a memory task. Results are in line with a developmental prediction. The percentages of children making realistic memory span judgments were greater in the two grade school groups; the percentages of children achieving correct recall when they indicated readiness were also higher for these two groups. The results were most clear-cut for the fourth grade group, and the data suggest that the accuracy of a child's "cognitive relation to his own memory system does somehow change in the course of the early school years"

(p. 332). Of equal interest is the fact that memory monitoring processes exist at the early grade school stage.

Hart proposed that a MEMO process leads to a FOK experience. However, the nature of this intervening process was not specified. Hochhaus (1970, Study I) related FOK accuracy to editing processes, and tested the hypothesis that partial recall is the basis for the FOK phenomenon. That is, although a S makes an unsuccessful recall attempt, he might have access to partial information (attributes) about the item (cf. Brown & McNeill, 1966). A FOK judgment may therefore be based on this implicitly retrieved information; the S edits the amount of partial recall, and then makes a decision concerning predicted recognition performance. In the experiment, nouns were paired with "highly codifiable" conceptual figures which were constructed of four specific dimensions (size, color, shape, & number). Each dimension was characterized by one of two values (e.g., large, small). After the study phase, Ss were to specify the correct value for each dimension, and then predict future recognition of the correct answer. Correct recognition performance was analyzed as a function of FOK judgments and partial recall scores (correct value selection for less than four dimensions). Both sets of data were accurate indicators of recognition performance, and it was concluded that knowing part of an item may be one explanation for the FOK phenomenon.

Adams (1967, 1968) cited the FOK and TOT data as being examples of response-monitoring behavior: the TOT/FOK phenomena may represent graded judgments of suppressed responses. When a person is attempting to recall a TOT target word, an awareness of the degree of closeness is used as a guide in the search process. The FOK data shows that Ss have an accurate feeling of their ability to make a correct response on a recognition test.

The discussion thus far has proceeded from a documentation of FOK/TOT experiences to studies of their accuracy in predicting future retrieval. A series of studies will now be presented which demonstrate that recall cues can increase the accessibility of blocked information. Unsuccessful recall of an item still in storage may be due to insufficient "retrieval rules" for that item. Thus, it is possible that some items which are initially not recalled may be made accessible by appropriate retrieval cues.

Freedman and Landauer (1966) presented a general information questionnaire to their Ss, who were instructed to give a certainty of knowledge rating from a 4-point scale to unanswered questions. In the next step of the procedure, the questionnaire was again presented, but this time a correct first letter cue was given for one-third of the items, an incorrect cue for one-third, and no cue for the remaining third. The final phase of the experiment was the administration of a 6-choice recognition test. A comparison of those items missed in phase 1 (initial questionnaire) but recognized in phase 3 showed a significant effect of confidence rating, indicating that Ss were able to give accurate graded judgments of knowledge about unanswered questions. The cue variable was also significant. When the correct initial letter was given in phase 2, .34 of previously unanswered items were correctly retrieved, compared to .12 and .15 successful retrieval for wrong clue and no clue items, respectively. This result may be compared with the Brown & McNeill (1966) finding that the correct initial letter of TOT items was often retrieved by Ss; it also confirms the results of the Horowitz et al. (1963) cue experiment.

Further evidence for the beneficial effect of first letter clues comes from a study by Hopkins and Atkinson (1968). In a series of four experiments, Ss were to give (a) the names of people represented in pictures (Experiments I & II), or (b) authors of books (III & IV). In I, the clue was given if a S could not name the person, whereas in the other three studies, the first-letter clue was given immediately before presentation of the stimulus. Each experiment also contained a within-Ss condition in which no clue was given. In all four studies, a small, but statistically reliable, clue effect was found. That is, more correct recalls resulted when the first letter of the target word was provided.

Data from other sub-conditions in Experiments I and II confirmed results from previously cited studies. In II, letter cues were sometimes given which could be either correct or incorrect. Ss had to judge the correctness of the clue, and did so significantly better than chance, confirming the Brown & McNeill (1966) finding that first letter knowledge is sometimes accessible. In I, Ss were to rate immediately upon presentation of the pictorial stimulus whether they would respond with the correct name. High ratings generally were followed by a correct response, whereas low ratings resulted in recall failure, even though first-letter clues were given. Thus, "... the confidence rating data indicated that S was able to judge to his potential recall performance with considerable accuracy, providing further empirical support for the notion that S can monitor the contents of his memory..." (p. 856).

A "judgments-of-knowledge" study by DaPolito et al. (1968) provides some evidence as to how Ss are able to recognize the correct answer from a set of alternatives. Again, a general information questionnaire was administered to the Ss, and they were to make a Yes/No recognition prediction after each recall failure. Ss then were given a 5-choice (homogeneous alternatives) recognition test for each failed item, and were told to point out those alternatives which they felt to be definitely incorrect; finally, Ss chose what they felt was the correct alternative. Of those items judged Yes, .58 were recognized correctly, whereas .32 No items were correctly identified, a result well in line with the Hart data. The key comparison was the proportion of incorrect alternatives eliminated as a function of Yes/No judgment; significantly more wrong choices were eliminated for Yes items than for No items. The fact that Ss can selectively eliminate wrong items suggests that they compare their partial information with cues from the recognition test, and decide whether or not a match has occurred.

The TOT/FOK data indicate that Ss are able to evaluate cognitive states. This fact is of some importance:

> The demonstration that people can rely upon subjective experiences to predict the contents of their memories is of general interest to psychologists concerned about the functional relationships between subjective processes and behavior processes (Hart, 1967b, p. 196).

Furthermore, just as Ss are able to make accurate FOK ratings about blocked information, so too are they able to give accurate confidence ratings concerning the correctness of their overt responses. In a memory signal detection theory (SDT) paradigm, a S is required not only to produce a response, but also to judge it. The Type II d' is an index based on the rating data (Clarke, Birdsall, & Tanner, 1959), and is a measure of S's ability to discriminate the "rightness" and "wrongness" of his responses. The ratings themselves are assumed to be independent of the response (Bernbach, 1967), and as such are a measure of one's knowledge of response correctness. Each of the theories of recall discussed earlier contains a covert editing process which judges whether an implicitly retrieved item is a to-be-recalled item. In the absence of overt recall, prediction about recognition performance (FOK), or a feeling that recall is imminent (TOT), would seem to be a result of a monitoring process which is sensitive to retrieved attributes or item-images of a target. If search recovers no information relevant to a query, then it is unlikely that a TOT or FOK state would result. An editing mechanism also seems to account for a subject's ability to rate overt-response correctness (see e.g., Hochhaus, 1969).

Research Problem

The present experiment was designed to investigate further the retrieval mechanism of recall. Although retrieval is a covert process, an experimental monitoring of it is afforded by a technique called pupillometry, in which pupil size of the eye is the dependent variable. Very briefly, this technique provides a continuous measure of "mental effort" during a cognitive task. That mental effort is involved in retrieval of information would seem to follow from the theoretical descriptions of this memory component: its stages are proposed to be governed by the control processes of the S (Shiffrin & Atkinson, 1969, e.g.), which include biases and strategies, initiation of search, and decision-making.

The purpose of the study was to determine whether the pupil response would be able to differentiate various states of knowledge as Ss attempted to answer general information questions. Overt responses to a question could fit one of three general categories: (1) an answer is given (it is either correct or incorrect); (2) a statement is made that the answer is not known; (3) a statement is made that one feels he knows the answer, but for the moment it is blocked. Since we are able to decide rapidly what information we do or do not have available, the level of processing to a given query would seem to depend on the intensity of the search, the number of attributes recovered, and the judging of such information for possible output. It was hypothesized that the measure of mental effort could distinguish between questions one answered (successful retrieval), questions he did not know (information not available), and questions he felt he knew (a knowledge of availability, but information presently inaccessible).

In order to study more directly the search process, a clue trial followed those questions which were not answered (correctly). The first and last letter of the answer was presented, and the nature of the search represented locating not only an answer of a given category (e.g., an author), but also one which fit the constraints of the clues. Thus, one's task was essentially that of problem solving, to construct a plausible response on the basis of partial information. The goal was to determine whether the pupil would respond differentially to clue trials which resulted in a correct answer as opposed to those trials which resulted in no answer.

Finally, an answer trial followed unsuccessful retrieval to the clue trial. The correct answer was given, and the S was to rate whether he now realized the connection between the question and the answer. It was hypothesized that those trials resulting in a positive link between question and answer would show a different pupil response than those in which the answer meant little to the S in regard to the question. The phenomenon represents the familiar "Of course!" or "That's it!" response to an answer we have been trying to recall, and finally achieve (possibly with a little help). A common example is trying to remember someone's name. Incorrect alternatives are quickly rejected, but when the right name is either stumbled upon or presented to us, we now know that we know the answer and are out of our momentary "misery".

Because the major dependent variable is pupil size, a detailed discussion on pupillometry will now follow.

CHAPTER II

PUPILLOMETRY

Psychophysiology and Pupillometry

Pupillometry is a psychophysiological technique. The dependent variable is physiological data (pupil size), whereas psychological stimuli serve as the independent variable. This relationship is to be distinguished from physiological psychology: "The latter deals with the manipulation of physiological variables and the recording of behavioral events while psychophysiology deals with the manipulation of behavioral events and the recording of physiological variables" (Stern, 1964, p. 90). The physiological data is used to make inferences about psychological phenomena.

Pupillometric data is easily obtained. The technique generally involves procedures originally reported by Hess (1965). A 16-mm motion picture camera is mounted to one side of a box-like structure. The subject looks through an opening at one end and fixates on a central point at the far end (generally a mark on a screen). Illumination is held constant. A mirror inside the box reflects an image of _S_'s left or right eye into the lens system. Typically, filming occurs at either 1 or 2 frames per second. The majority of pupillary studies to be reported herein involved the use of auditory stimuli; a few reports which used visual presentation (slides projected onto the rear screen) will also be discussed. Pupil sizes are obtained by projecting the processed film and measuring horizontal pupil diameter. The statistical treatment of the data will be considered in an upcoming section.

The basic notion behind the use of pupillometry in the studies to be considered is that continuous measures are obtained which reflect the amount of mental effort to a task. The first reports on dilations (increase in pupil size from some reference point) during mental tasks interpreted such responses as indicators of "mental activity" (Hess & Polt, 1964) or momentary cognitive load on the S (Beatty & Kahneman, 1966; Kahneman & Beatty, 1966). Further research has validated these interpretations, and the issue will be discussed more thoroughly in another section. In general, however, interpretations of physiological responses to behavioral events should be placed in their proper perspectives. The "mind-body" issue is of relevance to pupillometry as well as any other psychophysiological technique, and will be discussed next.

A series of stimuli, say a list of digits, is given to an S, and on the basis of pupil dilations at particular points in time, inferences are made concerning such hypothetical constructs as mental effort, cognitive load, or rehearsal. Most of the resulting descriptive language, therefore, deals with psychological, rather than physiological, terms. This is not to say, however, that the experimental events could not be described in physiological terms. Behavioral language is used because, for the purposes of the psychologist, it is more convenient and better suited to his needs. Also, the physiological vocabulary for a construct such as mental effort is just not "rich" enough at the present time. The usage here of "psychological vocabulary", and "physiological vocabulary" is based on the concept, "Linguistic Parallelism" (Graham, 1967). This notion is one proposed solution to the mind-body problem, and as Graham

(1967, 1971) has tried to show, current thinking on the issue is improper and confused; because the problem is prevalent in the fields of medicine, psychosomatics, and psychophysiology, a correct usage of terminology is important.

How the issue relates to pupillometry may be seen by considering the events in a typical memory task. A set of stimuli (digits) causes a response (pupil dilation). Now, to talk about a distinct psychological event (e.g., mental effort) and a physiological event is misleading, according to Graham. That is, to talk as if there were a dual reality (mind and body) is incorrect. "The duality is, however, not of the events observed; it is of the language used to describe them" (1967, p. 59). Thus, linguistic parallelism refers to the fact that the organism functions as a whole -- one aspect -- but that there are two ways of describing the behavior. For example,

> ... an emotion is some collection of events in the organism; we can give it a name like 'fear' or 'anger' which are words in the psychological language, or, we can use the names of processes in the nervous system, glands, and muscles, names which are words in the physical language" (Graham, 1971, p. 123).

Depending upon the researcher's goals and objectives, one language is chosen over the other in order to facilitate study of a specific problem area.

A key point in this parallelism-of-languages notion is the goal of translating one language into the other: "The mind-body problem, then, may be said to be the question of how the physical and psychological languages are related to each other" (Graham, 1967, p. 56). Ax (1964) uses an analogy to describe this translation: a symbolic program (e.g., Fortran) interacts with a (physical) computer by means of a translator which converts commands into machine language (and later reconverts to the symbolic language). The relationship between the psyche and soma is also a symbolic one; just as the symbolic computer program is as real as the machinery, so too is the organism's symbolic system -- the psyche -- real, and describing the symbolic relationship is what the facts and principles of psychophysiology attempt to achieve.

Psychologists who use pupillometry attempt to translate between the two languages in that a measure of autonomic activity is discussed in psychological terms. Different mental tasks are hypothesized to induce different cognitive operations (psychological states) which in turn have physiological concomitants in the form of differentiable patterns of pupillary responses. The term psychological "state" is a widely used one and deserves discussion. According to the arguments developed by Graham (1971), it would be improper to talk of a psychological state as one distinct aspect which causes a physiological state -- another aspect; ramifications of the mind-body problem would arise. The question of what caused a pupil response should be directed at the stimuli or task given to a S. The pupillary data may be used, however, to infer the properties of the proposed intervening process which operates between the stimulus and response. The main goal is to develop the psychological language such that a rich cognitive vocabulary develops in relation to our intervening variables.

Pupillometry offers a means to invoke the concept of "converging operations", (Garner, Hake, & Eriksen, 1956) which are "any set of two or more experimental operations which allow the selection or elimination of alternative hypotheses or concepts which could explain an experimental result" (pp. 150-151). The gathering of pupillary data is a way to achieve "methodological triangulation" in that it provides a second dependent measure on which to base inferences concerning the properties of a given psychological process. Stoyva and Kamiya (1968) discussed how the combining of verbal report with physiological data was instrumental in furthering the development of the hypothetical construct dreaming. They also emphasized the point that validation of other concepts in the study of consciousness could result from the use of "convergent indicators".

A number of pupillary studies have concerned themselves with determining what the pupil response measures. It is important to be able to specify whether a "common denominator" is being "tapped", because as Kahneman, Peavler, & Onuska (1968) have pointed out, "A variable that measures everything measures nothing very well ... Interpretation of pupillary changes as indicators of processing load are suspect when al ternate interpretations are available" (p. 187). In a section dealing with this issue, it will be seen how various experimental operations have converged on alternate explanations in an attempt to validate the pupil response as an index of mental effort or processing load. Once the pupillary measure has itself been validated, the technique may then be used as a convergent indicator on issues in other research areas.

In order that the pupil response to behavioral events be understood, a neurophysiological description will now be given. A discussion of the statistical treatment of pupillary data will then follow; the last section of this chapter will discuss what psychological statements and implications may be obtained from pupillary studies.

The Neurophysiology of Pupillary Activity

The size of the pupillary aperture is controlled by expanding or contracting movements of the iris, which itself is controlled by the interplay of muscles innervated by the autonomic nervous system. Parasympathetic fibers connect the iris sphincter muscle, whereas sympathetic fibers innervate the dilator muscle (Lowenstein & Loewenfeld, 1964). The light reflex (constriction of the pupil when the eye is exposed to light) is due to the parasympathetic reflex arc; efferent (motor) impulses travel from the oculomotor (Westphal-Edinger) nucleus to the sphincter. The response to light enhances visual acuity and thus allows vision over a wide range of light levels. The average latency of the response to light has been measured at 0.18 second, while the time from stimulus onset to maximal contraction usually takes 1-2 seconds (Adler, 1959). Pupillary diameter at maximal contraction is about 2 mm (under normal conditions, the diameter is between 3 and 4 mm).

Other reflexes resulting in pupillary constriction are the near reflex (accommodation convergence reaction) and the lid-closure reflex (Davson, 1963). The constriction due to the near reflex serves to control depth of focus:

> When the eye is focused for distant objects, the depth of focus is large and the pupil-size is relatively unimportant; when the eye is focused for near objects, on the other hand, the depth of focus becomes small and a constricted pupil contributes materially (Davson, 1963, p. 283).

Dilation of the pupil occurs in response to a decreasing level of illumination. Maximum diameter of the pupil may reach 8 mm. Dilation also takes place to affective or psychic stimuli:

> Under the influence of sensory or emotional stimuli, afferent impulses reach the central nervous system. Such excitation, or, in man at least, spontaneous thoughts or emotions, give

> rise to descending discharges from cortex and thalamus which activate the great sympathetic centers in the hypothalamus. ... Simultaneously with the active sympathetic discharge to the dilator of the pupil, inhibitory impulses converge from the cerebral centers upon the oculomotor nucleus and prevent if from sending constrictor messages to the pupillary sphincter (Lowenstein & Loewenfeld, 1964, p. 145).

Dilation due to direct sympathetic discharge is more rapid than that due to oculomotor inhibition. As will be seen shortly, dilation also results to psychological stimuli in tasks where "cognitive effort" is demanded. The interpretation of this phenomenon on a neurological basis is an interplay of the cortico-thalamic-hypothalamic mechanisms.

Pupil size in any given stimulus situation depends on the "dynamic equilibrium" between parasympathetic and sympathetic tonus on the corresponding iris muscles. Under normal conditions, continuous shifts in this equilibrium cause "pupil unrest" -- small fluctuations in pupil size. When this unrest is severe, the condition is called "hippus".

Statistical Treatment of Pupillary Data

How a response is defined in psychophysiological research is to a large extent dependent upon the investigator. Lacey (1956) attempted to attain some degree of uniformity in measurement in an article entitled "The evaluation of autonomic responses: Toward a general solution". In recent years, articles concerning measurement and statistical issues have frequently appeared in the journals Psychosomatic Medicine and Psychophysiology. A discussion of measures of autonomic nervous system functioning is provided by Heath and Oken (1965). The diversity in definitions of a response is not limited to the more classical dependent variables of, say, heart rate (HR) or palmar conductance. As will be seen, many different measures, as well as statistical tests of significance, have been employed in pupillometric research.

The most "popular" form of pupillary data presentation is the "response level" method. That is, unadjusted absolute values of pupil size (in millimeters) are presented as a function of time (e.g., Kahneman & Beatty, 1966). A less frequent means of defining pupillary responses is the "relative" technique of expressing the response as a change from one kind of baseline; reactivity to a stimulus is defined in terms of magnitude of change. One method is to include a baseline period on every trial, and to calculate the deviation of the actual size at each point in time from the mean of the baseline (e.g., Stanners & Headley, 1972). The baseline is assumed to represent S's pupil size at "rest". Any fluctuations during the baseline period may be due either to "biological noise" (pupil unrest) or to random thoughts. Change scores have also been produced by adjusting the scores to a common baseline at the first second (e.g., Kahneman, Peavler & Onuska, 1968, Experiment II). Some form of a change score is necessary for statistical purposes when the pupil response is being compared with other autonomic measures, such as the galvanic skin response (e.g., Kahneman, Tursky, Shapiro, & Crider, 1969). Another baseline method is to film a control and response period on each trial, and then treat these two periods as a variable to be analyzed statistically (e.g., Simpson & Climan, 1971). This procedure is different from other baseline methods in that absolute scores are used; pupillary reaction to a stimulus is compared to pupil size prior to the onset of the stimulus (control period).

A measure which is frequently used in visual-presentation methodologies is the percentage of change from the pre-stimulus (control) period. Less frequently used measures include the latency of response (time from onset of stimulus to maximum dilation), and the dilation rate/unit of time or stimulus item. Nontemporal statistics include the magnitude of peak dilation, and the mean response level (average of unadjusted scores over the total response period).

Although all pupillary studies provide the investigator with pupillary responses as a function of time, not all experimenters choose to present and/or analyze the time dimension; the abscissa, when used, is presented in units of frames, seconds, or "time blocks". One advantage of a psychophysiological technique is that it provides time-locked, continuous measures such that response patterns to different stimuli can be followed over time. Time-series analyses, however, present certain statistical "problems". Typically, repeated-measures designs are used in pupillometric research: tasks are repeated on the same Ss. The time dimension must also be considered a repeated measure in that a given number of values are collected on each trial. However, pupillary measures gathered on any single trial cannot be assumed to be independent of each other. The value at point x in time is determined in part by the value reached at the preceding point. This dependence is represented statistically by correlation -- "Among repeated measurements taken in time, there frequently is serial correlation present: the measurements taken closer together in time are more highly correlated than those taken further apart" (Danford, Hughes, & McNee, 1960). Serial correlation in time series can not be avoided since time can not be randomized -- that is, second 5, say, must always follow second 4, and precede second 6 (Hence, time is a fixed variable). It might also be noted that were it not for serial correlation in psychophysiological measures, characteristic patterns of responding as a function of time would not be revealed. For example, in some of the pupillary studies to be discussed, a linearly increasing "loading" phase is found during stimulus (digit) presentation, followed by a gradual decrease over time as the stimuli are reported ("unloading"); such a uniform pattern would not occur if a present value were not to some extent dependent upon the preceding one.

As an example of such correlation in pupillary measures, data for a single S was taken from a study by Stanners & Headley (1972). Four treatments were randomly presented 10 times each to Ss; one of these conditions (1) showed greater overall dilation than the others, whereas condition 4 showed the least overall dilation. The average pupil response to the 10 trials for each of these two conditions was used to calculate serial correlation coefficients (SCC), according to procedures outlined by Holtzman (1963, pp. 200-205). A SCC of lag 1 pairs the $n+1^{th}$ value in time with the n^{th}, the $n+2^{nd}$ value with the $n+1^{th}$, etc; likewise, a SCC for lag 2 pairs the $n+2^{nd}$ value with the n^{th}, and a lag of 3 pairs the $n+3^{rd}$ value with the n^{th}. Product-moment correlations (r) are then performed on the resulting pairs of observations for each lag. A significant r (t test; null hypothesis, correlation = 0) implies a sequential-dependency relationship for a given lag; that is, to a large extent, the pairs of observations vary in unison. The first 10 lags were computed for both conditions mentioned above. The "correlograms", along with the pupillary response curves, are given in Appendix A; correlation is graphed as a function of lag. In condition 1, serial correlation is present in all 10 lags, whereas in condition 4, only the first two lags are significant. It is apparent that even in data having little serial correlation, sequential dependency does exist for those points closer together in time (lags 1 and 2), reflecting the fact that a given value is partly dependent on the previous value.

A related issue to serial correlation is the non-homogeneity found in the various components of error in the repeated-measures analysis. It is generally the case in psychophysiological data that the subjects X time error term is smaller than the resulting subjects X stimulus conditions (treatments) error term. The reason for such a difference, as noted by Graham (1970, pp. 486-487), is due to the higher correlation of those observations closer together in time. This higher correlation results in a smaller variance. Graham (1970) illustrated this point for HR data by comparing error terms for time with those for treatments over several studies: in every case, the time error term was smaller. This pattern is not limited to HR data. From analyses of variance tables (unpublished) gathered from Stanners, Headley and Clark (1972), Stanners and Headley (1972), and Clark (1970), comparisons were made of error terms from pupillary data. Thirty-four different tables were examined, and within each table, the Ss X treatments error terms (for main effects) were divided by the Ss X time error term. A total of 66 such comparisons revealed that the treatments error terms were on the average 7.4 times larger than the time error terms (in all comparisons, the former term was larger; the range was 1.2 to 21.8 times larger, and the standard deviation was 16.4; 24 comparisons resulted in values greater than 10). Although the smaller error variance for time may be partly due to the relatively larger number of degrees of freedom (df) accumulated, the trend is still likely to exist even when the df are the same for both error terms: 4 of the 66 comparisons had equal df, and the treatments error term was larger by an average of 7.2.

The main point of the above discussion is that separate components of error should be used to test the appropriate variables. Graham (1970) has shown that different results can occur when the sources of variance are pooled into a common error term. Such a procedure is inappropriate because the pooled term will be too small to test for treatment effects (the Ss X time component contributes the most df), resulting in an overestimation, and will be too large to test for the time effects, resulting in an underestimation (Graham, 1970, p. 486). A check of pupillometric studies which included a temporal variable in the analysis showed that the data were analyzed according to procedures commonly outlined for repeated measures designs (e.g., Winer, 1962, Chap. 7). Because the error variance is separated into various components, overestimation of effects is less likely to occur, and any bias would seem to be in the negative direction (Type II error). This partioning procedure makes the resulting F tests more conservative than those tested under the pooled-error-term procedure.

An analysis which includes a time dimension affords many meaningful comparisons. A significant effect of time in a pupillary study indicates that a consistent response pattern was obtained. A significant treatments effect implies that the amount of pupillary activity was not the same for all tasks. A significant treatments X time interaction indicates that distinct response patterns for the tasks occurred during the course of the trial. Based on such information, inferences may be made concerning the nature of the cognitive processing involved in each task.

A related point concerning the use of F tests centers on the preceding serial correlation discussion. The symmetrical matrix assumption that error of measurement is independent from one observation in time to the next is probably violated (the assumption specifies equal variances and covariances for the x values of time). In order to take into account possible violations of this assumption, a more conservative test may be initiated by reducing the df (e.g., Myers, 1966, pp. 160-162). This procedure has been applied by a few authors (e.g., Simpson & Molloy, 1971; Stanners, et al., 1972; Clark, 1970). Again, it should be noted that the time variable can not be randomized, and that serial correlation in psychophysiological data is a naturally occurring function of neurophysiological mechanisms; a change of the pupillary aperture from a basal size of, say, 3.0 mm to 3.5 mm is time dependent.

Although the time variable is generally graphed by most researchers, it is not always analyzed. A more widely used statistical test than the $\underline{F}$ test is the $\underline{t}$ test with data averaged over time periods. Since the data is time-locked to stimulus events, the resulting pupillometric curves may be closely scrutinized. For example, an investigator may wish to test select phases of curves, rather than the means of the whole response period. Or, one may desire to test the difference between the maximum values, or latencies, of two tasks.

The following considerations, as noted by Graham and Clifton (1966) also make a time plot seem advantageous. If the response period is averaged, specific autonomic activity to a particular stimulus event will not be detected. Furthermore, a change in direction may not be noticed:

> If, for example, HR increased during the first 5 seconds following a stimulus and then decreased below baseline in the next 5 seconds, the average HR for 10 seconds would reveal no change or would reflect only whichever component, acceleration or deceleration, was larger (p. 307).

Nonparametric statistics have been used to test whether magnitude of response orders itself as a function of proposed task difficulty. More frequently, it is used to test the percentage change measure. This measure however, may be somewhat inadequate: as the initial (basal) level of a measure increases, a given amount of change receives a pro-ressively smaller value. For example, if three Ss have resting heart rate levels of 65, 75, and 85 beats per minute, and each shows a 10 unit increase to the same stimulus, the resulting percent changes would be 15%, 13%, and 12%, respectively (see, e.g., Sternbach, 1966, pp. 45-46). As an example from the pupillometric literature, Schaefer, Ferguson, Klein, and Rawson (1968) presented the percent changes which occurred to a multiplication task. Pupil diameter increases of .5 mm for two different Ss resulted in percent increases of 8% and 11%. Likewise, two changes of 1.0 mm were given respective change values of 17% and 25%, two 1.5 mm changes resulted in values of 27% and 30%, and two 1.8 mm changes resulted in values of 35% and 43%. In each case, the higher percent change score was due to the lower prestimulus value.

An important question related to the issue of change is whether a 10 unit increase in a heart rate measure, say, has the same (psychological) significance regardless of the "resting" level before stimulus presentation. Much research has been conducted on this issue, and data suggests that for many autonomic measures, the prestimulus level of functioning should be taken into account. The reason is described by the law of initial value (LIV), first examined by Wilder (1958). Briefly, the LIV

> ... relates the response to a stimulus with the prestimulus level of the responding system. The law states that the higher the prestimulus level (initial value), the smaller the tendency to rise with exciting stimuli, and the greater the tendency to drop following inhibiting stimuli; and with extreme high or low levels there is progressive tendency toward no response or for reversal in the direction of response (Oken & Heath, 1963, p. 3).

Hence, a 10 unit increase from a high basal level may "mean" more than a similar increase from a lower level. Investigators working with the more traditional autonomic measures of heart rate and palmar conductance have been concerned with the LIV issue (the journal Basimetry deals solely with this problem), and various statistical procedures have been proposed to handle autonomic data (e.g., Lacey, 1956; Sternbach, 1966, Chap. 4; Wilson, 1967). The LIV phenomenon may be a function of homeostatic processes: as a measure is already near its limits, negative feedback mechanisms attempt to oppose any further changes.

Whether the LIV pertains to pupillometric research should be examined, because pupillary activity is under the control of the dynamic equilibrium of certain effector mechanisms (Lowenstein & Lowewenfeld, 1964). Although the dilations generally observed to mental tasks are relatively small compared to responses which are invoked by the light reflex (a mean increase of 1.0 mm to a task would probably be a record), the LIV might still be a consideration, given the fact that basal levels vary widely due to different laboratory procedures.

One physical parameter which probably has the greatest effect on basal level in the laboratory is the illumination level used for photographic purposes. Bradshaw (1969) specifically tested the effect of "baseline levels within a wide but nonextreme range of illumination" on the pupil response to a reaction time (RT) task. The major finding was that while the lower light level produced higher baselines, the shape of the two curves showed little difference in form or magnitude. The effect of different basal levels due to drug state was later tested by Bradshaw (1970) on the same RT task. Three drug conditions were used -- alcohol, amphetamine, and normal. The latter two conditions are of interest here. The amphetamine condition resulted in a higher baseline than the normal, but the shape and amplitude characteristics of the curves for these two conditions did not differ.

It would appear, then, that the responding to the RT tasks carried the same significance regardless of the basal level. Somewhat of a similar note is provided by several authors who studied the effect of the pupil response to an imagery task. Correlations between prestimulus levels and amount of dilation were not significant in three studies reporting such a comparison (Colman & Paivio, 1969, p. 297, fn,; Simpson, 1969, p. 118, fn.; Simpson & Molloy, 1971, p. 495, fn.); it should be noted, though, that the tasks were ones which evoked relatively small changes in pupil size.

Depending on the task, however, illumination levels might be a factor influencing the response pattern. Kahneman, Beatty, and Pollack (1967) had their Ss engage in a combination digit transformation-signal detection task. Illumination was found to affect rate of constriction during the latter portion of the response curve in that higher illumination resulted in more constriction during the last 2 seconds. The authors interpreted this result in the following manner: high illumination causes the light reflex mechanisms to constrict the pupil; as soon as mental effort to a task subsides, the antagonistic effect on constriction is removed, and a more rapid decrease in pupil size occurs. Wright and Kahneman (1971) applied this reasoning to a verbal behavior study in which sentences were either to be memorized or understood. Their goal was to obtain information "... about the distribution of mental effort within sentences, and about the relative difficulty of various parts of a sentence... (p. 198). A low level of illumination was used in hopes of detecting mental effort (although the basal level would be relatively higher, there would be no competing mechanisms to constrict the pupil), whereas a high level was used to detect any cessation of mental effort during sentential processing. However, there was no interaction of illumination level with response patterns, so the light variable was discarded, and the data combined.

Complete implications of the LIV for pupillometric research are not as yet completely spelled out. The evidence presented above indicates that responses to a task are similar even when prestimulus levels are different (but not near the limits of 2 to 8 mm). Some investigators have noted the basal issue and have dealt with it in some manner. Fredericks and Groves (1971) used a predetermined "modal level of illumination" in their study of the pupil response to emotional stimuli presented pictorially. This level gave the pupil an equal chance to dilate or constrict. Electronic pupillometric equipment which provides on-line monitoring of actual pupil size allows the E to control basal size by adjusting illumination levels; such a system is described in Schaefer et al. (1968).

The notion of the LIV is important in interpreting a phenomenon which sometimes occurs in pupillometric research (as well as other techniques which measure autonomic responses) -- habituation, or adaptation, of the response to similar stimuli. Such decreases over trials have been noted to stimuli ranging from simple sentences (Stanners et al., 1972) to digits (Kahneman & Beatty, 1966) to multiplication problems (Hess & Polt, 1964), and to verbal stimuli of different affective value (Lehr & Bergum, 1966). Often there is an accompanying "down-drift" in baselines over the course of the experiment, as Bradshaw (1967) and Kahneman and Beatty (1967) have reported in their data; the preceding discussion on the LIV can rule out the decrease in responsiveness as being due to the baseline changes -- according to the law, lower initial values would result in larger increases to a stimulus which generally evokes a dilation. An important question, then, is whether the lability (reactivity) of the pupil decreases with repeated exposure because of fatigue or boredom with the task ("Form, extent, and duration of pupillary reflexes are influenced by the degree of tiredness, wakefulness, or excitement of the subject..., Lowenstein & Loewenfeld, 1964). The light reflex habituates to repeated stimuli (Adler, 1959, p. 181), and hence it may be the case that the response system fatigues in mental tasks as well. According to discussions on this issue by several authors, the habituation can be attributed to cognitive factors: e.g., "reduced psychological significance of the stimulus" (Kahneman & Beatty, 1967), adoption of a "consistent performance set" (Kahneman & Beatty, 1966), or a "learning-to-learn effect" (Stanners et al., 1972). The issue is succinctly summarized by Kahneman and Peavler, 1969: "The adaptation of autonomic responses during continued exposure to a task is a psychologically significant result rather than a characteristic of the response system" (p. 315).

A number of points follow from this discussion of pupillary adaptation and basal changes. First of all, it might be meaningful to show that certain kinds of tasks are susceptible to progressive decreases in response whereas other tasks are not. Also, if several different tasks are to be used in a study, a significant Time X Tasks X Periods interaction (where Periods represents the data grouped in trial blocks) would imply that the initial pattern and/or magnitude of responding is differentially affected by repeated experience with the tasks. Generally, however, investigators want to control for habituation effects. In this case, different tasks should be presented randomly in order "to randomize arousal increment and decrement effects across stimuli" (Woodmansee, 1966, p. 134).

Furthermore, in light of possible down-drifts in basal size, it would not seem appropriate to calculate deviations from a predetermined value measured prior to the experiment per se. When change scores are to be determined, the use of a baseline period on every trial would compensate for habituation effects. In the sense that a S serves as his own control in a repeated measures design (each S receives all treatments and his "noise" affects all trials), so too does the baseline-on-every trial procedure serve as a control against a progressive decline in responding.

Finally, it is important to consider baselines of between-Ss groups when absolute scores are used (approximately 40% of 34 pupillary studies checked had a between-Ss variable). If random placement, or different illumination levels, resulted in one group having a substantially larger baseline than the other, problems could arise in interpretation of results, because one group would show consistently larger values than the other. (This issue is of particular importance if one is dealing with tasks which evoke relatively small dilations.) Wright and Kahneman (1971) noted that the prestimulus levels were somewhat different between their two groups, and in order to make a reliable test of the between-group treatment effects, they converted their absolute score data to change score measures.

To summarize this section on the statistical treatment of pupillary data, it would seem wise for an investigator to consider the importance of controlling for basal changes over the course of an experiment. Also to be considered is the potential information which is available from the inclusion of a time dimension in an analysis. Furthermore, one should be aware of certain statistical "problems" in analyzing psychophysiological data, and should be aware of the reasons behind such procedures as using component error terms, and reducing the degrees of freedom in statistical tests.

What Does the Pupillary Response Measure?

Pupil responses have occurred to a wide assortment of mental tasks, and have been interpreted in psychological language as reflecting cognitive effort. Such an interpretation arose from some earlier studies, and further investigations over the years have for the most part supported this explanation. This section will discuss the early studies, and then will trace the validation of the pupillary technique as a viable measure of cognitive processes.

In 1964, Hess and Polt reported a relationship between amount of pupil dilation and the proposed difficulty of multiplication problems. For a given problem, the general response pattern was a gradual dilation which reached a peak just before S responded; post-solution pupil size showed a gradual decrease. Bradshaw (1967) found similar relationships of phasing and magnitude to arithmetic problems of varying difficulty. Such response patterns were found to occur to other tasks as well. Beatty and Kahneman (1966) extended the pupillometric technique to a study of short-term memory (STM) and long-term memory (LTM) for telephone numbers. The dilation pattern for an unfamiliar number (STM) was a gradual increase which peaked at the start of a report phase, then decreased with report. Recall attempts for previously learned numbers (LTM items) resulted in a more rapid dilation which reached a greater peak, but showed a similar report phase pattern. These authors in another study termed the dilation portion of the curves the "loading" phase, and the constriction portion the "unloading" phase (Kahneman & Beatty, 1966); in this study a variety of STM tasks were used, and previous results were substantiated. Also, magnitude of response was found to be a function of task difficulty.

These two 1966 studies were a beginning in establishing the pupillary response as a measure of mental activity. Both within task as well as between task validity was suggestive: within a trial, second-by-second activity "... provides a very effective index of the momentary load on a subject as he performs a mental task" (Kahneman & Beatty, 1966, p. 1584). Between tasks, the extent of this load is reflected by the magnitude of dilation.

Further support for the notion that the pupil response is an index of cognitive activity is supplied by studies which show a strong relationship between behavioral data and autonomic activity. That is, both sets of data are found to be mutually supportive of each other. As one example, data from a divided-attention study by Kahneman, Beatty, and Pollack (1967) shows a relationship between pupil response and performance on a signal detection and/or digit transformation task. While a S was attentive for the presence of a "K" in a display which flahsed letters at the rate of 5/second, he also was required to perform a 1-unit transformation on a 4-digit string (e.g., 8340 transformed to 9451). For comparative purposes, some trials required only the transformation tasks, while others involved just the detection task. A trial lasted 8 seconds, and consisted of listening, pause, and response phases; a K could occur at any time during the trial. In the double-task condition, detection performance was highly contingent on the point in time of presentation of the K. That is, detection progressively decreased as the digits were being presented and processed, then progressively increased as the answers were being reported. The pupillary response pattern showed a strikingly suggestive relationship with detection performance: the pupil dilated during digit presentation (indicating digit processing), but decreased in size during the report phase (a decreasing processing load). At times when digit processing was least (lower pupil size), detection performance was best (hence the interpretation was that the pupil is sensitive to momentary fluctuations in processing load -- an example of internal validation). External validation was suggested by the fact that performance was best in the detection-only condition, and pupil dilation was smallest to this task (the pupil response was thus reflecting task difficulty).

Both kinds of validation were supplied in another study dealing with sensory discrimination (Kahneman & Beatty, 1967). Ss heard a 850 cps standard tone, and judged it to a comparison tone (CT) which ranged from 820 to 880 cps. Although the dilation patterns were similar for a 880 cps CT (easy discrimination) and a 850 cps CT (hard discrimination), magnitude of response was greater to the harder discrimination. The most substantial pupillary activity occurred during a 2 second period immediately following presentation of the CT; to be noted is that the average of this period varied as a function of tonal frequency of the CT -- dilation was greatest near 850 cps, but became progressively smaller as the CT deviated on either side of this value. Percent errors followed the dilation pattern very closely. Performance was worse near 850 cps, but became progressively better as the CT was shifted towards 820 or 880 cps.

Mutually supportive evidence is suggested by a substantial series of studies dealing with the pupillary response to imagery tasks. Paivio (1966) found that abstract words required more time than concrete words to elicit images, and it was reasoned that the former task was a more difficult one. The hypothesis that the pupillary response would reflect this difference in cognitive effort was then tested; dilation was found to be greater to abstract than concrete words (Paivio & Simpson, 1966; Simpson & Paivio, 1966), apparently indicating the greater difficulty in forming images to abstract nouns. However, a more sensitive measure than magnitude of dilation proved to be latency of dilation (time from onset of noun to peak pupil size). The maximum value occurred sooner to concrete than abstract words in one study (Paivio & Simpson, 1968), and in another (Simpson, Molloy, Hale, & Climan, 1968), latency differentiated between nouns rated either high, medium, or low in imagery value. Finally, latency of dilation was found to be in accord with key press latency (Colman & Paivio, 1969). Ss pressed a key when an image was formed, and this latency as well as pupil latency was shorter for concrete nouns.

The above experiments show a relationship between behavioral and pupillary data. The following studies show a correspondence between pupillary changes and the postulated nature of covert processes. Pupillometry is undoubtably a better method than introspection to study certain constructs: inferences concerning their characteristics may be based on the notion that processing load or mental effort is referenced by the pupil response.

One construct which has been a "natural" subject for pupillometric assessment is that of rehearsal, a process which occurs as a S actively attempts to memorize a set of stimuli. The intensity of rehearsal should be reflected in the amount of mental effort expended on the task, and hence pupil size should be an indicator of various rehearsal strategies.

Kahneman, Onuska, and Wolman (1968) pursued the implications of the response pattern of the Kahneman and Beatty (1966) study. Presumably, the steady, progressive increase in pupil size during stimulus presentation reflected an increase in the rate and amount of rehearsal activity. Dilation was greatest at the beginning of the report phase, a time when rehearsal activity should be the most active. The decrease during report may be due to a decreasing processing load. The 1968 study was designed to test the pupil's sensitivity to various kinds of rehearsal activity. Citing research which suggests that rehearsal is "cumulative and repetitive" for ungrouped stimuli, but "intermittent and non-repetitive" for grouped stimuli, the authors presented a series of 9 digits to their Ss; the strings were either given at a 1 digit/second rate, or were presented in three groups of three digits each. The pupil response to the ungrouped string showed a progressively increasing pattern, whereas in the other condition, brief increases followed each group's presentation, and a substantial increase occurred after the last group, suggesting a "pulling-together" effect. Although serial learning was required in both tasks, the pupillary response patterns differed. Thus, the measures provided an accurate index of the nature and intensity of rehearsal during the trial.

The linear increasing function which occurs for ordered recall does not occur in a free recall paradigm. Kahneman and Peavler (1969, sub-experiment) presented a list of 8 nouns at a 1 word/4 second rate. The presentation of each noun was followed by a dilation-constriction pattern, and the pupil size at the end of the trial was lower than at the beginning. It appears that rehearsal was of a different nature in this free recall situation.

Rehearsal strategy as a function of task demands was studied by Kahneman and Wright (1971). Three category groups of 4 items each were presented on a trial; Ss were instructed before each trial began as to whether they would be required to repeat all 12 items (whole recall) or just one group (probed recall), and whether the retention interval would be 3 or 7 seconds. Depending upon the memory task and the retention interval, the authors reasoned that rehearsal strategies would differ in accord with the task demands, and that the pupil response would reflect these differences. The predictions were for the most part confirmed, in that systematic pupil changes occurred to the different tasks.

Finally, a study by Johnson (1971) provides further evidence for the notion that the pupil response is sensitive to alterations in processing strategy. Using an "intentional forgetting" paradigm, Johnson presented lists of 5 nouns in a serial recall task. Half of the lists contained a signal cue which indicated that all the words preceding the cue were not to be included for recall. The cue occurred at four different serial positions. The response pattern for the non-cue trials showed the typical loading and unloading functions of Kahneman and Beatty (1966). However, the pupil response to the cue trials showed a systematic dilation-then-constriction pattern immediately following the cue. The dilation portion was interpreted as an attempt on Ss part to categorize the list into to-be-remembered and to-be-forgotten subsets. The constriction phase, which lasted for 2-3 seconds, suggested a reduction in processing requirements, in that pre-cue items no longer needed to be rehearsed. These results were in line with predictions based on intentional forgetting theory, and provided support for that theory, and the pupil measure as well.

A concept which is related to rehearsal and processing strategy is the notion of instructional set: if a S knows beforehand what the task requirements are for a given trial, will he process a common set of stimuli differentially to meet different demands? This question was investigated by Stanners and Headley (1972), who presented a randomized series of four different tasks. Every trial included a list of 5 digits, followed 6 seconds later by a probe-recognition test. The task requirements differed in accord with the combination of whether a rehearsal-interference activity was required immediately after digit presentation, and whether recall of the digits was required after the recognition test. Differences in pupil response occurred during a Pause phase (between the pretask instructions and the first digit), and the Digit phase. The largest dilations were in response to the most difficult task (intervening activity plus recall), whereas the smallest dilations were to the least difficult task (probe-recognition only). Pupil size to the other two tasks were similar to each other and were situated between the most and least difficult tasks. Of some note is the fact that the above ordering of tasks occurred during both the Pause and Digit phases, suggesting that a processing strategy is preceded by a preparatory state.

Similar results were obtained in a study by Clark (1970) in which recall/no recall was a between-Ss variable. Ss were presented auditorially a list of digits, followed a few seconds later by a probe digit. A cue just before the probe instructed the S whether or not he would have to decide if the digit was a member of the preceding sequence. One group of Ss was informed that they would have to recall the digits on every trial, while another group was not given this memory requirement. It was found that significantly greater dilation occurred during digit presentation by those Ss who had to recall. Also, when Ss were instructed that a probe decision was required, greater dilation occurred following the cue as compared to the trials when Ss were informed that no probe decision was necessary.

The effect of no pretask information on digit processing is shown by Peavler (1969, Experiment I), who presented a randomized series of digit lists which varied in length (5, 9, or 13 digits). The response patterns to the early portion of the three list-types were identical; differences in magnitude did not become apparent until points were reached which identified the trial as either short, medium, or long. Presumably, if Ss were told the trial-type previous to its presentation, differences in patterning at the outset might be expected.

Colman and Paivio (1970, Experiment II) studied the pupil response to mediational processes in paired-associate (PA) learning. They were interested in determining whether pupil size to the task would differ under instructional sets of imaginal mediation, verbal mediation, and no mediational instruction. Ss were given four study and four test trials on a list of 16 noun-pairs; average pupil size showed a general decrease over study trials for the two mediational sets, whereas it maintained a somewhat constant size for the no-set condition. Although recall performance was equal by the fourth test trial for the three groups, the pupillary data suggested that the task was harder when no mediational instructions were given -- pupil size was thus more sensitive to the cognitive demands than was recall performance.

Kahneman and Peavler (1969) used a paired-associate task to study the effect of incentive on pupil size. Specifically, these authors hoped to be able to predict from the pupil response during study trials which items -- high-reward (HR) or low-reward (LR) -- would be successfully recalled. A list of 8 digit-noun pairs was presented on eight study-test trials, and on each trial, half of the pairs were given a HR value. Recall performance averaged over trials was better for HR items (55% vs. 18%). The pupillary response pattern to study trials for the two incentive groups showed a similar dilation to the digit, but greater pupil size occurred to the noun under the HR condition, indicating a greater effort to learn the pairs associated with a HR value. On test trials, significantly less dilation occurred to LR blanks (no overt response given) than to HR blanks or HR correct trials; this finding suggests that retrieval processes were not initiated when the stimulus item of a LR pair was presented, since learning of such pairs was minimal.

Studies on the pupil response to attentional processes show that pupil size prior to overt responding is able to differentiate the actual responses eventually given by a S. The divided attention study by Kahneman et al. (1967), and the sensory discrimination study by Kahneman and Beatty (1967) have already been discussed. Studies on vigilance were conducted by Hakerem and Sutton (1966). In one study, the task was to report whether or not a threshold-level light was seen on a given trial. A constriction to the stimulus (light reflex) occurred only when the light was reported as seen. In a follow-up study which used a stimulus of insufficient light energy to cause a contraction, but of sufficient energy such that it would be seen 50% of the time, a pronounced dilation occurred only on those trials to which S responded with "seen". The pupil response to "not seen" trials was no different from trials on which no discrimination was required, and also was no different from those which had a stimulus of insufficient energy to be seen. "It seems that the mechanism which controls the response of the pupil to light stimuli is related to the mechanism which controls the threshold for seeing" (p. 486). For dilation to occur under these stimulus conditions, the light must not only be seen, but must also have significance: no diltion occurred to the trials on which no discrimination was required; hence dilation to a detection (hit) is a response to more than the stimulus per se.

Finally, pupillometry appears to be a useful tool to study certain cognitive notions of psycholinguistic theory. A recent topic of investigation has been the study of the "psychological reality" of syntactic structures: that is, do the rules prescribed by a grammar have a psychological counterpart. Inferences concerning the cognitive representation of such structures are often made on the basis of recall test performance. However, some investigators have used paraphrasing tasks. Fillenbaum (1971) questioned whether "indirect" testing procedures such as recall are "... primarily directed toward the assessment of the psychological reality of some syntactic structure or, say, toward an analysis of how a sentence might be learned or remembered" (1971, pp. 276-277). The question, then, is whether the processing of sentences in tasks requiring understanding might not be different from the way in which material is processed for a memory task. To study this issue of whether the tasks "tap" the same structures, Stanners et al. (1972) monitored the pupil response both during and after presentation of simple sentences. One group of Ss was required to memorize each sentence, whereas another group was instructed to understand the sentence. Dilation during a post-sentence phase was greater for the paraphrase condition, indicating that a more complete processing of the material may have occurred for this task. The implication is that the cognitive representation of sentential material may be a function of the task.

Wright and Kahneman (1971) also studied the effect of sentence processing on pupil size. On the hypothesis that strategies differ according to the task, they informed their Ss that either recall of a complex sentence, or understanding of it, would be required on a particular trial. Systematic differences in pupil size occurred during a post-sentence retention interval of both 3 and 7 seconds. The authors concluded that there are alternative processing strategies of sentences for retention purposes. To be noted is that each of these two pupillary studies provided a convergent indicator to test alternate explanations of sentential performance data.

The major interpretation of the pupillary studies cited in this section is that increases in pupil size correspond to increases in mental effort or processing load. However, other interpretations have been proposed. The pupil dilates to any sympathetic stimulation (e.g., Adler, 1959; Lowenstein & Loewenfeld, 1964), and some investigators have questioned whether the notion of mental effort best accounts for responses to mental tasks. Perhaps better explanatory terms would be anxiety or arousal -- that is, pupil size may be measuring differences in arousal due to different task demands. Other questions have arisen on methodological grounds; these issues will be discussed first, followed by a discussion on the arousal hypothesis.

To begin with, it must be shown that changes in pupil size during a mental task are not a result of sensory stimulation per se. Nunnally, Knott, Duchnowski, and Parker (1967) found that the pupil response to tonal stimulation is related to the intensity of the tones, in that the loudest tone produced the largest response. It may be the case, therefore, that changes are partially due to sensory effects. In the Peavler (1969) study, control trials were run to test this notion. Lists of digits (5, 9, or 13) were presented to Ss who were instructed simply to listen to each list. No significant dilations occurred and it was concluded that sensory stimulation is probably not a confounding factor in pupillary studies dealing with cognitive tasks.

The effect of Ss' prior knowledge of pupillary studies was investigated by Clark and Johnson (1970). Specifically, they were interested in whether the loading and unloading phases characteristically obtained in STM studies might be partially due to Ss' attempts to produce desired responses. Ss were given a serial recall task involving 6 nouns. Different instructional sets were induced by informing one group of Ss about the dilation-constriction pattern normally observed in such tasks, by misinforming another group about an expected constriction-dilation pattern, and by not giving a third group any information concerning pupil responses to mental tasks. Results showed the typical loading and unloading phases to the task for all three groups, with no significant differences between them. Hence, knowledge of pupillary studies does not appear to be a factor.

Some investigators have studied the possibility that changes in accommodative convergence are responsible for certain pupillary changes during mental tasks. If the fixation point is not at a distance such that a S can maintain his focus for the duration of a trial, accomodative shifts could have a constrictive effect. The issue is worth considering, because a check on the fixation distances used by investigators shows a variety of values ranging from 6 inches to 8 feet.

Hess and Polt (1964), using multiplication problems, found only slight differences in mean increase in pupil size for fixation distances of 3 1/4 feet, and 10 1/4 feet. Kahneman and Beatty (1966) used fixation distances of 6 inches and 6 feet in their STM tasks. Although the far point produced a larger baseline, and hence larger overall values, the pattern of the loading and unloading phases were very similar under both fixation conditions. Furthermore, because a dilation phase did occur to the near point fixation (where constrictive effects would be more likely -- see the earlier neurophysiological discussion), and a constriction phase did occur to the far point fixation (where the accomodation factor should be minimal), it appears that the accomodative effects, if present, do not override the experimental effect on pupillary changes under constant illumination conditions.

Pupillometry is not the only autonomic-measuring technique which is sensitive to changes in mental effort. Kahneman et al. (1969) gave digit transformation tasks of three levels of difficulty to their Ss. Pupil size, heart rate (HR) and skin resistance measures were collected. All three measures showed systematic changes during processing and report phases. Also, the three measures were able to differentiate the effort involved in the transformation tasks. Colman and Paivio (1969) measured pupil size and skin resistance to an imagery task involving concrete and abstract nouns. Although only pupillary magnitude differentiated the two classes of nouns, latency of both the pupillary and GSR measures made the distinction. The fact that the HR and GSR measures show changes to mental tasks lends more credence to the notion that pupil size is an autonomic index of mental effort.

A major consideration in interpreting pupil responses is determining what effects certain overt task requirements have on pupil size. Specifically, most tasks required either a manual (e.g., key press) or a verbal response. One question which then arose was whether pupil dilations occurred as a result of anticipation of a motor act, an interpretation which would seriously affect the mental effort notion. Some of the earlier studies reported different magnitude of responding due to response requirements. Bradshaw (1967) noted that verbalization at specified points during arithmetic and anagram tasks resulted in a larger pupil size. Hakerem and Sutton (1966) found larger overall pupil size to a vigilance task when a verbal report was required after each trial. Simpson and Paivio (1966) attempted a replication of their first imagery study (Paivio and Simpson, 1966) to determine what effect a key-pressing response had on pupil size. (Ss indicated occurrence of an image by pressing a key.) In this second study, Ss merely attempted image formation for a specified period of time; results showed an attenuation of the pupil response, and a lower level of discrimination of abstract nouns from concrete nouns based on the pupil response.

Several later studies also found various forms of overt responding responsible for increased pupil size. Bernick and Oberlander (1968) reported that verbalizing one's though sequences produced greater dilation than a no-verbalization condition. Nunally et al. (1967) obtained systematic increases or decreases in pupil size as Ss were required to lift weights of increasing or decreasing magnitude, suggesting a relationship between pupil size and muscle tension.

It would appear that dilation occurring to mental tasks is to some degree confounded with overt response requirements. Various explanations have been proposed to account for the effect, and the more salient are as summarized by Simpson and Molloy (1971, pp. 491-492) and Kahneman et al. (1968, pp. 193-194): 1) an overt response requirement may affect the S's strategy in performing the task, and hence more mental effort may be expended to a task which includes some form of a response; 2) the anticipation of making an overt response leads to increased activation (muscle tension); 3) an overt response causes anxiety about experimenter evaluation; 4) organization of material for response purposes may increase processing demands. Explanations 1 and 4 may be considered cognitive interpretations, whereas 2 and 3 involve variables not directly related to processing load. Studies will now be cited which have dealt with these issues, and it will be seen that, although overt requirements do affect the pupil response to some degree, the critical variable is the processing requirements of the tasks themselves.

Simpson, Paivio, and their co-workers have provided the most substantial testing of the possible ways in which overt response requirements may affect the pupil response to various mental tasks. In Paivio and Simpson's 1966 study, the requirement to press a key upon image formation may have created a more demanding task than one in which a S attempted to form an image during a specified time period (Simpson & Paivio, 1966). That is, the tasks may not in fact be the same tasks; when a key press is required, S must explicitly decide whether or not an (acceptable) image is present, and this decision process may add to the amount of mental effort. To test this possibility, Simpson and Hale (1969) had Ss engage in a decision-making task; one group was required to decide which of two given directions they would move a lever upon a signal; a control group was told which direction to move the lever. Even though both groups were required to make a manual response, the experimental group showed greater dilation during the period preceding the signal, implying that a relatively easy decision still results in a larger pupil size. Thus, in certain tasks an overt response may require a S to attend to the task more fully.

The effect of an overt response on pupil size appears to depend on whether the response is "task-relevant" or not. One group of Ss in a study by Simpson and Paivio (1968) were cued to begin a description of their image to a word 10 seconds after the stimulus was presented; Ss in another group were instructed to form an image, but to recite the alphabet when cued. Significant dilation following the word occurred only in the former group. The impending description task apparently had an effect on the effort evoked, since less dilation occurred when the explicit response was not related to the task.

Similar results occurred when the relevant/irrelevant response was a motor task. Simpson (1969) required Ss in one group to indicate their judgment to a pitch discrimination task by a lever movement; Ss in another group were also required to make the same judgment, but lever movement was unrelated to the task. Relatively little dilation occurred during the discrimination period in the irrelevant condition, whereas larger changes took place in the condition requiring an explicit indication of discrimination.

Simpson and Climan (1971) measured both pupil and electromyographic (EMG) changes to their usual imagery task (control period, noun presentation, image period) to determine whether an increase in pupil size corresponded with an increase in EMG activity in anticipation of, or in the act of, making a manual response. Five different groups were used, and were distinguished according to key-pressing instructions. Based on the interpretations of inter-group comparisons, it was concluded that muscle activity per se does not directly affect pupil size. For example, in a group which was required to key press from the time of image formation to the end of the image period, EMG activity remained high, but pupil size decreased during the latter portion of the period. Another group was required to press the key throughout the control (prestimulus) period, whereas yet another group was told to press it throughout the image period; similar dilations were observed in the first half of the image period of both groups, even though the key pressing occurred in different periods. Furthermore, although EMG activity maintained itself throughout the respective key-pressing period, pupil size decreased during the latter half of the period.

A comparison of those groups performing a task-relevant key press with those performing a task-irrelevant press indicated that the relevant press resulted in a greater magnitude of pupil change to the task, confirming earlier results. Thus, muscle activity itself is not the sole determiner of the pupil response, since task relevancy is a factor; the fact that pupil size decreased while EMG activity remained at an increased level implies that the pupil response is more sensitive to the second-by-second processing efforts of a S than to methodological artifacts.

The possibility also exists that the requirements of an overt response causes anxiety due to concern about experimenter evaluation. In order to test whether such anxiety affects pupillary responding, Simpson and Molloy (1971) took Ss who scored either high or low on the Audience Sensitivity Inventory (ASI), and gave them digit transformation tasks of either high (1-unit transformation) or low (repeat original digit-string) difficulty. The crucial comparison was the response pattern of the ASI groups during a 3 second Pause period, which occurred between the Listen and Report phases of the trial (report involved responding in front of two experimenters). Pupil size was larger for the high ASI group throughout this Pause period, and it remained at a constant level, whereas the response pattern of the low ASI group showed a decline over time. During the Listen period, however, only the task variable showed a significant effect (the more difficult task elicited the greater dilation in both ASI groups). Thus, it may be inferred that the task manipulation itself was mainly responsible for differences during the processing phase, although emotional factors may have influenced the pupil response to some additional degree.

Kahneman, Peavler, and Onuska (1968, Experiment I) used these same transformation tasks in a study which manipulated a verbalization requirement. The purpose was to determine what confounding effect this variable might have on pupillary activity during mental tasks. Task difficulty was manipulated by combining the size of the transformation with a "say" or "think" requirement: the former instruction meant that after the listen phase, the answer should be verbalized, then immediately repeated; the latter instruction meant that the answer should first be covertly thought, then verbalized. Initial vocalization for either task resulted in a relatively small increase in pupil size, but was then followed by a consistent decrease during the remainder of the trial. The pupil responses to the tasks were given a cognitive interpretation by the authors. Because task difficulty was the major influence on pupil size, the effect of the overt response requirement was assumed to have an organizational effect on processing requirements. This organizational interpretation is also in line with the Beatty and Kahneman (1966) and Kahneman and Beatty (1966) studies in that peak responding occurred about 1 second into the report period (followed by a consistent decrease), suggesting a rapid organization of the material for response purposes.

Another alternate explanation for pupil responses to cognitive tasks is the arousal hypothesis, which states that emotional or motivational influences occur in proportion to the task difficulty. That is, dilations indicate task anxiety; harder tasks, or those given a higher incentive value, may be accompanied by greater anxiety, and the sympathetic arousal is reflected in pupil size. The notion of "task arousal" is different from the autonomic activity which is assumed to accompany deliberate mental effort on the part of the S: "Arousal is often constructed as an essentially automatic reaction to significant or overwhelming stimuli, whereas processing load refers to the demands imposed by activities in which S engages, often voluntarily" (Kahneman & Peavler, 1969, p. 317).

The arousal interpretation has been used to explain some pupillary data. In the Nunnally et al. (1967) study, one sub-experiment involved the anticipation of a gunshot. Ss were told that the numbers 1 through 5 would be presented, and that the shot would occur sometime during the presentation of number 3. Pupil size increased during numbers 1 and 2, reached a peak at 3 (actually the gun was never fired), then decreased during numbers 4 and 5. The authors interpreted these pupillary responses, and those from their other tasks, as indicating "a general measure of activation". Bradshaw (1967) assumed the intervening variable in his mental tasks to be level of arousal. Also, there is a substantial body of literature concerned with the pupillary response to affective stimuli (reviews of this area may be found in Goldwater, 1972; Hess, 1968). Thus, it may be that arousal is a better explanatory concept of pupil size in mental tasks than, say, processing load or cognitive effort. In an attempt to clarify the issue, data from relevant studies will be discussed next.

Experiment II of the Kahneman et al. (1968) study was designed to determine what effect a motivational variable (incentive) would have on the course of pupillary changes to transformation tasks. Different conditions were created by the combining of size of transformation with size of a monetary reward for correct performance. Results indicated that this incentive manipulation had only a slight effect on response patterns, and that task difficulty was the more powerful variable. The incentive manipulation influenced only the high-incentive/easy-task condition; dilation was greater during the listen phase as compared to the low-incentive counterpart. However, this result was interpreted as indicating a greater level of mental effort. Although failure was more probable in the two hard-task conditions, incentive did not affect them, suggesting that the difference in the easy-task conditions was due to increased effort rather than arousal.

Polt (1970) presented two different groups of Ss two series of multiplication problems. Prior to the second series, the experimental group was informed that a shock would occur with each incorrect answer. There was no difference in pupil response between groups on the first series of problems, but on the second series, the experimental group showed larger dilations to the tasks. This latter result was given a cognitive interpretation; it was argued that the larger size was the result of increased expenditure of mental effort due to the threat of shock (none was actually given). Because pupil size during the pre-task period was the same for both the "shock" and control groups, it would seem that the larger increase of the shock group to the task was a result of a greater effort to solve the problem correctly, and was not therefore due to a "fear" response (which is the interpretation of the Nunnally et al. (1967) threat-of-gun-shot experiment).

In the Kahneman and Peavler (1969) study previously mentioned, the reward value to a PA task was varied. The task involved digit-noun pairs, and Ss were instructed that either an odd or even digit signified either a high-reward (HR) or low-reward (LR) pair. Dilation to the digit on study trials did not vary as a function of incentive; differences did show when the noun was presented, indicating that at this time greater effort was exerted to HR pairs. If emotionality were directly affecting pupil size, then one would expect differences to the stimulus, which served as an incentive cue.

Further evidence against the emotionality interpretation is provided by the Peavler (1969) study. It will be recalled that Ss were presented digit lists which varied in size (5, 9, or 13 digits); they were not told beforehand the length of any given list. The response pattern to the 13 digit list showed a leveling effect around the ninth digit. If anxiety affected pupil size during mental tasks, then the realization that the trial was a "long" one should have resulted in at least a brief dilation, perhaps signifying frustration with processing overload. However, the response pattern was similar to that for 9 digit lists, implying that storage capacity was being reached and that processing was then confined to those digits already in storage.

The results of these last four studies imply that a purely arousal interpretation is not feasible. Kahneman and Beatty (1967, p. 104) suggest to those who wish to attribute all pupillary changes to an emotional interpretation that the concept of anxiety may either have to be made synonymous with "processing load", or else broadened so wide to accompany all the data that the concept might become meaningless. On the other hand, the evidence does not allow one to state that pupillary changes to mental tasks are produced exclusively by cognitive effort. The effect of an overt response, for instance, must be taken into account. Certain aspects of pupillary data might be best explained by an arousal interpretation. For example, Johnson (1971) felt that some portions of the data to his cued-forgetting study were due to emotional components. Thus, while the experimental effects of processing load are the more powerful variables affecting pupillary activity, some of the responding may be partially due to nonprocessing factors. When properly used, though, pupillometry provides continuous records of cognitive activity during all phases of a task.

CHAPTER III

METHODOLOGY

Subjects

Twenty-one subjects were obtained from undergraduate psychology classes at Oklahoma State University. They were given a few extra points toward their final course grade as an inducement for participation. The data of nine subjects were excluded from analysis for two reasons: (1) three subjects were unable to maintain a steady fixation, making it impossible to obtain clear photographs of their eyes; (2) six subjects' responses to the questionnaire resulted in an overloading of certain response-categories, thereby leaving some cells unfilled. The film of the remaining twelve subjects was processed; there were seven males and five females in this group.

The following restrictions were required of the Ss: that they (1) have at least 20/30 vision without the aid of glasses or contact lenses, and (2) possess eyes that are light in color (e.g., blue, green). Restriction (1) was to ensure that the S would be able to fixate properly and comfortably on a distant fixation point, and restriction (2) was for photographic purposes -- light-colored irises provide a more distinct pupil-to-iris contrast on film, and thereby allow a more accurate measurement of pupil diameter.

Apparatus

The basic equipment used to obtain the pupillometric records consisted of a pupillometer and a 16 mm. motion picture camera. The pupillometer was a rectangular wooden box with dimensions of 22 1/2" x 22 1/2" x 48 1/2" (see Hess, 1965). The front end was equipped with chin rest and stationary eyepiece. The back end consisted of a rear projection screen (polyethylene covering) with a fixation cross (3/4" high; 1/2" arms) positioned in the center. The inside was painted flat black.

A Beaulieu R16 movie camera was mounted on the right side of the pupillometer; the camera was equipped with a Vemar 135 mm. f/2.8 telephoto lens, a Vemar "C" mount adapter, and 30 mm. of extension tubing. Camera speed was calibrated to 2 frames/second (exposure duration of .2 sec./frame), and to ensure constant speed throughout the experiment, the camera's separate power supply was connected to a voltage stabilizer (Raytheon VR6114). The film was Kodak Double-X Negative, Type 7222.

A half-silvered mirror was situated inside the pupillometer at a 45 degree angle from the S's line of vision to the camera. This positioning allowed S a view of the rear of the box, and also allowed a reflected image of the right eye to strike the lens system.

The experiment took place in a large room with a normal level of lighting (ambient level of 100 ft-c at S's eye level when seated; windows were covered with aluminum foil in order to control for changes in external light levels). Illumination inside the pupillometer was provided by a Kodak Carousel projector with the lens removed. A circular pattern of light was projected onto the rear projection screen. Illumination at S's eye was approximately 20 ft-c.

Materials were presented over a tape recorder (Uher Royal de Luxe) equipped with headphones for the S. Connected with the tape recorder was a sound-operated relay (Grason-Stadler, Model E7300A-1) which controlled a frame marker. The onset of a stimulus activated a pinhole light source (mounted on inside of eyepiece) which served to identify those frames associated with the stimuli. Camera operation was controlled by a sound-operated relay on the tape recorder; a cue was placed on one channel of the tape, and a connection to the camera allowed for remote control start and stop functioning via this cue.

Questionnaire

A pool of general knowledge questions was obtained in several ways. Most were devised by browsing through reference sources (The Lincoln Library of Essential Information; Reader's Digest 1972 Almanac and Yearbook; The New York Times Encyclopedic Almanac, 1972; The World Book Encyclopedia). Other questions were taken from lists used by Hart (1965a) and Hopkins and Atkinson (1968, Experiment III). Various combinations of these questions were used in a pilot study (17 Ss); the goal was to develop a questionnaire which would result in a distribution of responses into the cells of the experimental design (explained below).

Forty questions were selected on the basis of the pretesting data. All the questions were short (average length = 6 words; average duration = 1.98 sec.), and had only one correct answer (see Appendix B). The questions dealt with cities (10), authors (11), states or countries (6), natural science (7), and other assorted topics (6). The questionnaire was randomly ordered with the exception of the following restrictions: (1) the level of difficulty was varied in order to prevent strings of easy or difficult questions; (2) two questions dealing with the same topic could not occur in a row; (3) two questions whose answers have the same first and last letters could not occur in a row (this restriction was for the Clue trials). The same ordering of the questions was used for all Ss. Five additional questions were included for use as practice trials.

A given question had two other trials associated with it: a Clue trial which presented the first and last letters of the answer, and an Answer trial. The questionnaire thus contained 120 potential trials (40 x 3). All materials were tape recorded.

Procedure

A S was first checked for uncorrected vision of at least 20/30 as determined by a Snellen Scale eye chart. He was then seated before the pupillometer, and tape recorded instructions (Appendix C) were presented to him over headphones. The experimenter briefly summarized the task, and then presented the practice trials.

The sequence of events was as follows: for each trial type (Question, Clue, or Answer), the onset of a white warning light served as a cue for the S to position his head in the chin rest and to fixate on the distant cross. A few seconds later the word "Ready" was heard, followed four seconds later by the onset of the stimulus. Five seconds after the stimulus, the word "Respond" occurred, and the S leaned back from the apparatus to make a response. Filming began with "Ready", and ended with "Respond". The S fixated on the cross during this time, and was instructed to refrain from blinking. He was told to do his thinking during the post-stimulus phase so that he would be ready to react fairly quickly after the word "Respond".

After a Question trial, either a written answer was attempted, or a rating was made via toggle switch movement. The S was encouraged to give an answer if he felt he had "... a good idea of what the correct answer is". All answers were one-word answers. If a question called for a person's name, only the last name was required. If he could not give an answer, the S made a "Know" rating if he felt he did know the answer, but for the moment could not quite recall it; otherwise, he made a "Do Not Know" rating. The experimenter monitored the rating by the onset of one of two indicator lights which were connected to the toggle switch.

The ensuing trial type depended upon the outcome to the Question trial. The possibilities are presented in flow-chart form in Appendix D. If the S's answer was correct, a green light came on, and the experimenter moved the tape to the next question. An incorrect answer was made known to the S by the onset of a red light; in this case, or if a rating was given, a copy of the question was placed in S's view on the front of the pupillometer. Then a Clue trial was presented. Four seconds after "Ready" the first and last letters of the answer were given auditorially, followed five seconds later by "Respond". The S leaned back from the apparatus, and either gave an answer or a closeness rating: "Closer" meant that the clues brought him to the verge of achieving the answer; "No Closer" indicated that he was no closer to the answer than before the clues.

A correct answer was followed by a new question. If the S gave an incorrect answer or a rating, an Answer trial was presented (if the question called for a person's name, the first name was given first, followed by the last name). Five seconds after the answer, the S gave a recognition rating: "Recognize" indicated that he now realized that the answer was the obvious response to the question (the "Oh yeah", or "Of course" experience); a "Not Recognize" rating indicated that the answer meant nothing to him in relation to the question. After this rating, the copy of the question was removed, and a new question was presented.

In order to identify the film record of one trial from another, two blank frames were exposed between trials.

To summarize briefly, a S's response to the Question trial placed it in one of four categories: Answer Correct, Answer Incorrect, "Know", "Do Not Know". His response to a Clue trial likewise fit one of four possibilities: Answer Correct, Answer Incorrect, "Closer", "Not Closer". A response to an Answer trial placed it in either a "Recognize" or a "Not Recognize" category.

Before the experiment began, Ss were informed of the nature of the questions, and hence had a general idea of what to expect in the way of queries. Also, the instructions emphasized that all the questions were legitimate ones (that is, they all had a correct answer) and that all clues or answers they heard would be correct. An index card listing the various ratings and corresponding switch movements was placed by the toggle switch to ensure that S's ratings were correctly indicated.

The entire session lasted approximately one and a half hours. Before a S was dismissed, he was asked not to discuss the experiment with his classmates.

Scoring of Film

The processed film was displayed on a microfilm reader (Xerox Microforms Reader Model 2240) which projected an image to slightly greater than ten times its actual size. Pupil diameter was measured frame by frame to the nearest millimeter. Some frames were not measurable because of blinks or eye movements. Such frames accounted for only 2% of the 19,904 total frames, and there were no appreciable differences in the percentage of unscoreable frames associated with each category. Also, the percentage of unscoreable baseline (prestimulus) frames was no different than the percentage for post-baseline frames.

The trials were graded without knowledge of their response category.

CHAPTER IV

RESULTS

The subjects responded to an average of 84 trials on the questionnaire (out of a possible 120 trials if no correct answers were achieved), 46% of which were Question trials, 32% Clue trials, and 22% Answer trials. Table I presents the frequencies for each response category.

The pupillary data was prepared for statistical analysis in the following manner. The eight frames immediately preceding the onset of a stimulus were baseline frames, and the average of these frames was subtracted from each post-baseline frame. All the trials within a given category for each S were then averaged on a frame-by-frame basis. These averages were the units for the analyses of variance and other statistical tests discussed below.

All time series were divided into two periods for separate analysis. A Listening phase included those frames associated with reception of the stimulus, and a Processing phase included the remaining frames. In the case of the Question trial, the first five post-baseline frames were associated with the Listening period (calculation of the duration of the questions showed an average of 3.9 frames; the fifth frame was added to allow for a full interpretation of the question, which averaged 6 words). The other nine frames were placed in the Processing phase. Clue presentation required an average of 2.3 frames; the first three frames were placed in the Listening phase, and the remaining nine in the Processing

TABLE I

NUMBER OF TRIALS IN EACH TREATMENT

S #	Question				Clue[1]		Answer	
	AC	AI	K	DK	AC	NC	R	NR
1	9	11	4	15	6	19	8	16
2	11	5	10	8	6	16	7	11
3	9	3	4	21	5	21	12	11
4	15	3	9	11	8	13	8	8
5	8	10	10	11	13	15	12	6
6	12	5	2	18	12	11	3	9
7	14	8	12	6	8	13	11	7
8	11	3	8	18	6	15	14	9
9	10	4	13	13	7	19	13	9
10	12	3	4	19	8	13	10	9
11	16	8	12	4	7	5	14	3
12	14	8	12	6	7	6	14	4
Sum	141	71	100	150	93	166	126	102
Mean	11.7	5.9	8.3	12.5	7.7	13.8	10.5	8.5
%	14.9	7.5	10.5	15.8	9.6	17.5	13.3	10.8

[1]Answer Incorrect and "Closer" categories were excluded from analysis. See text.

phase. The presentation of an answer took an average of 1.8 frames, and the first three were treated as Listening frames, and the other nine as Processing frames. In both the Clue and Answer trials, as in the Question trial, the extra frames in the Listening period was to allow for full interpretation of the stimulus.

In all statistical analyses, the 0.5 level was adopted as the minimum for an effect to be considered significant. All tests were within-subjects, and in order to account for possible violations of the symmetrical matrix assumption underlying the statistical model, the conservative procedure of reducing the degrees of freedom to 1 and n-1 for all analyses of variance was undertaken (Geisser & Greenhouse, 1958).

Only 19% of the Clue trials resulted in either an Answer Incorrect (AI) or a "Closer" (C) category. The average number of trials/S was 1.7 and 3.4, respectively (5 of the 12 Ss had 0 or 1 trials for AI, and for C). Because this few number of trials would provide a very unstable score for an autonomic measure, these two categories were excluded from analysis (This pattern of outcomes for the Clue trial was not peculiar to this experiment; in the pilot study (17 Ss), subjects gave an average of 1.5 incorrect answers, and 3.1 "Closer" ratings.). The average number of trials/S was 82 with the omissions.

Question Trial

Time-Series Analyses

In order to determine whether differences in pupil response occurred to Answer Correct (AC) and Answer Incorrect (AI) trials, an analysis of variance (AOV) was performed. Figure 1 shows the two treatments as a function of time. The curves are identical during the Listening phase

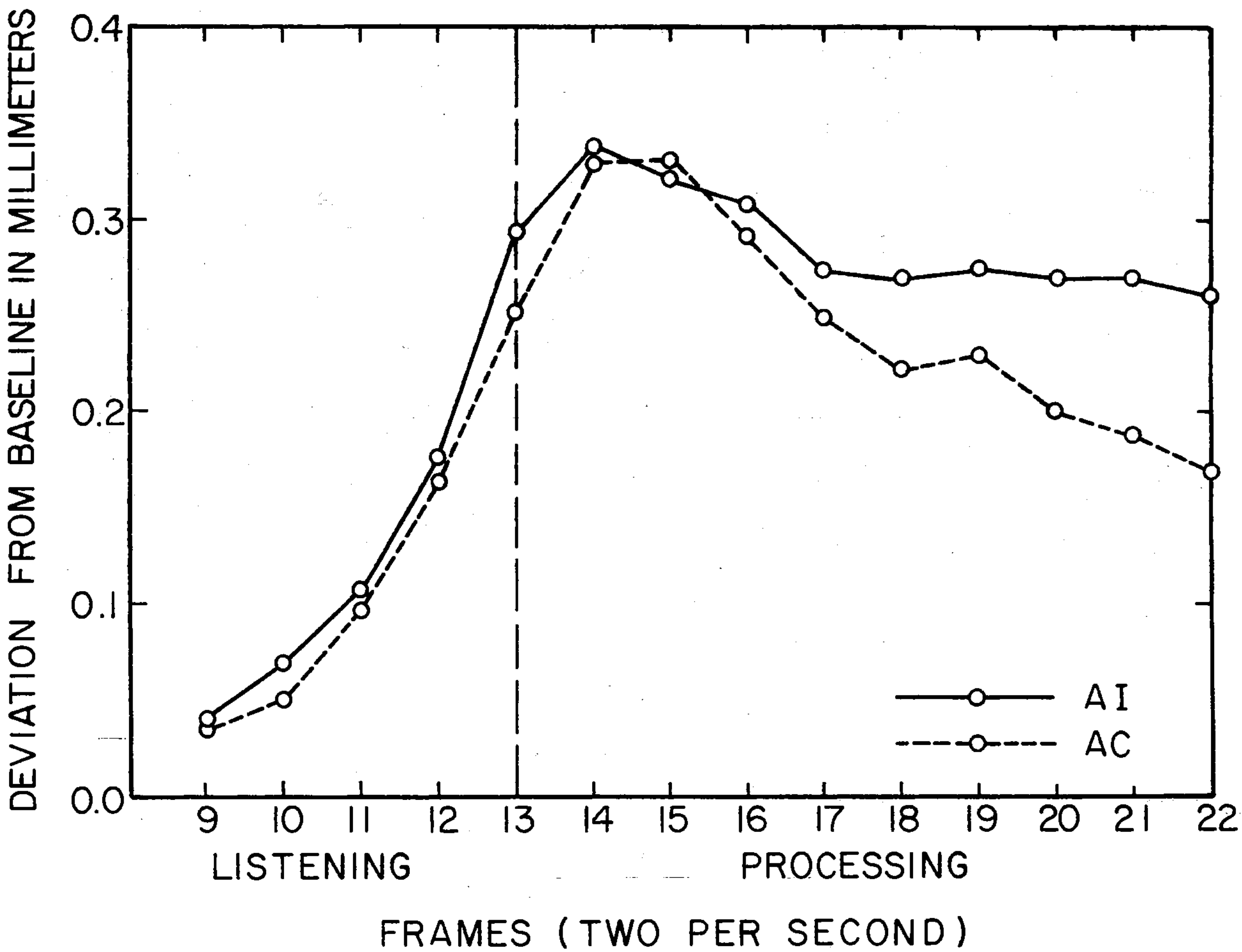


Figure 1. Pupil Response to Answer Incorrect (AI) and Answer Correct (AC) Outcomes of the Question Trial

($p > .20$). Although the treatments variable is not significant when analyzed over all frames of the Processing period ($p < .20$, Table II), the average of each treatment over the last two seconds (frames 19-22) shows a substantial difference in dilation (.269 mm vs .197 mm; $t = 2.263$, $p < .05$; degrees of freedom for all t tests reported herein = 11, and all tests are two-tailed). The difference between these last two seconds may be reflective of S's confidence in the correctness of his retrieval.

The two rating categories (Fig. 2) were also analyzed. There is no difference in the curves during the Listening period ($p > .20$). The "Know" (K) rating does diverge, however, from the "Do Not Know" (DK) rating during the Processing period ($p < .05$, Table III).

A procedural note is in order concerning the DK ratings. All the questions were devised on the assumption that Ss would be familiar with the key words in the question (e.g., that everyone would have heard of the Taj Mahal). It is conceivable that a differential pupil response might result to questions a person could not answer because of lack of knowledge (does not know the location of the Taj Mahal) as opposed to complete ignorance of the question (has never heard of the Taj Mahal). This possibility was considered after the fourth S was run; Ss 5-12 were presented with a list of key words from the questions, and were asked to check those items with which they were completely unfamiliar. This rating took place upon completion of the experiment. Since novel items accounted for only 5% of the DK ratings for these 8 Ss, it was not felt that the DK trials were confounded with an unwanted variable.

In order to evaluate differences in pupil size among the four treatments, Newman-Keuls procedures (Kirk, 1968, pp. 91-93) were performed on the means of the Processing phase (the treatments rose identically during

TABLE II

AOV OF PUPIL RESPONSE DURING PROCESSING PHASE TO AI AND AC OF QUESTION TRIAL

Source	df	MS	F (df corrected)[1]
Total	215		
Subjects (S)	11	0.39502	
Frames (F)	8	0.04526	8.65[d]
FS	88	0.00523	
Answer (A)	1	0.09151	2.83
AS	11	0.03231	
FA	8	0.00725	2.57
FAS	88	0.00282	

[1]See text for explanation. Note: significance levels are represented in all tables by the following: a = $p < .001$; b = $p < .005$; c = $p < .01$; d = $p < .025$; e = $p < .05$. Individual error terms (indented) were used in each F-ratio.

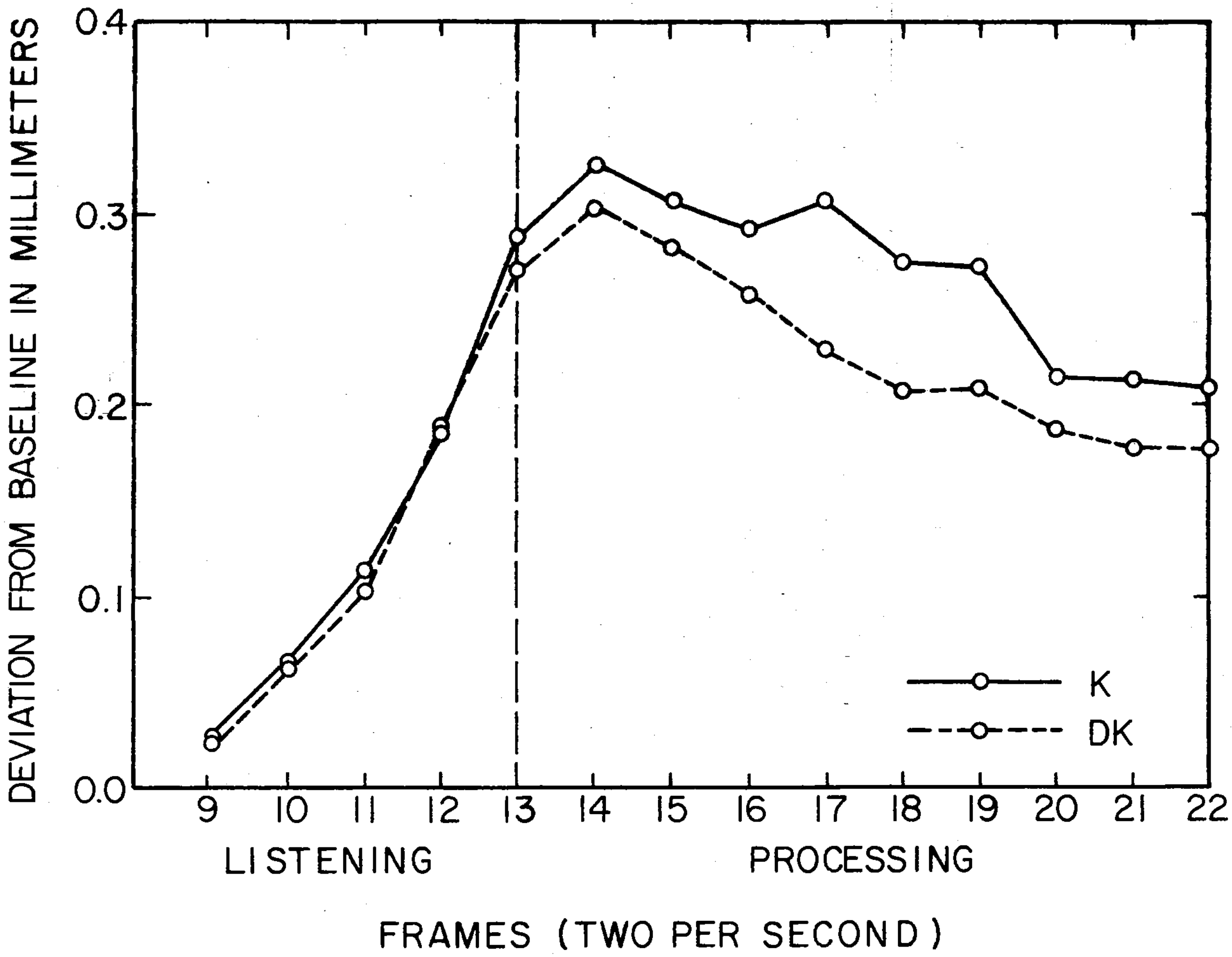


Figure 2. Pupil Response to "Know" (K) and "Do Not Know" (DK) Outcomes of the Question Trial

TABLE III

AOV OF PUPIL RESPONSE DURING PROCESSING PHASE TO K AND DK OF QUESTION TRIAL

Source	df	MS	F (df corrected)
Total	215		
Subjects (S)	11	0.37320	
Frames (F)	8	0.04591	14.13[b]
FS	88	0.00325	
Rating (R)	1	0.10170	6.49[e]
RS	11	0.01566	
FR	8	0.00266	1.61
FRS	88	0.00165	

the Listening period, $p > .20$). The means are represented in Figure 3 (dotted bars), and the curves are shown together in Figure 4 (a combination of Figs. 1 & 2). The F-ratio on treatments is suggestive of differences ($.05 < p < .10$, Table IV), but the multiple comparison tests revealed only a difference between AI and DK ($p = .05$). Certain aspects of the data did show differences, however, and these findings are presented for purposes of later discussion. Frame 15 of AC and DK is of different magnitude (.332 mm vs .282 mm, $t = 3.049$, $p < .02$). This difference will be alluded to in the Clue-trial results section. The averages of the last three frames for AI and K are different (.267 mm vs .215 mm, $t = 2.696$, $p < .05$), possibly indicating a difference between the weighing of a potential item (AI) and the decision on retrieved attributes of an item (K).

Peak Dilation and Latency

An alternative measure to overall responding is the maximum response to a treatment. The average peak change score was determined for each treatment/S and an AOV was performed on these units (Table V). The Treatments variable was significant ($p < .05$), and Newman-Keuls comparisons on the means (Fig. 3, striped bars) revealed that AI and K were each different from DK ($p = .05$).

A related measure to the peak response is the latency of response, defined here as the number of frames from the onset of the stimulus to the peak. This value was determined for the mean peak of each condition/S, and an AOV was then performed (Table VI). Although the Latency variable was not significant, ($.10 < p < .20$), a Newman-Keuls test on the mean latencies (AC: 6.83; AI: 7.92; K: 7.25; DK: 6.42) showed the AI

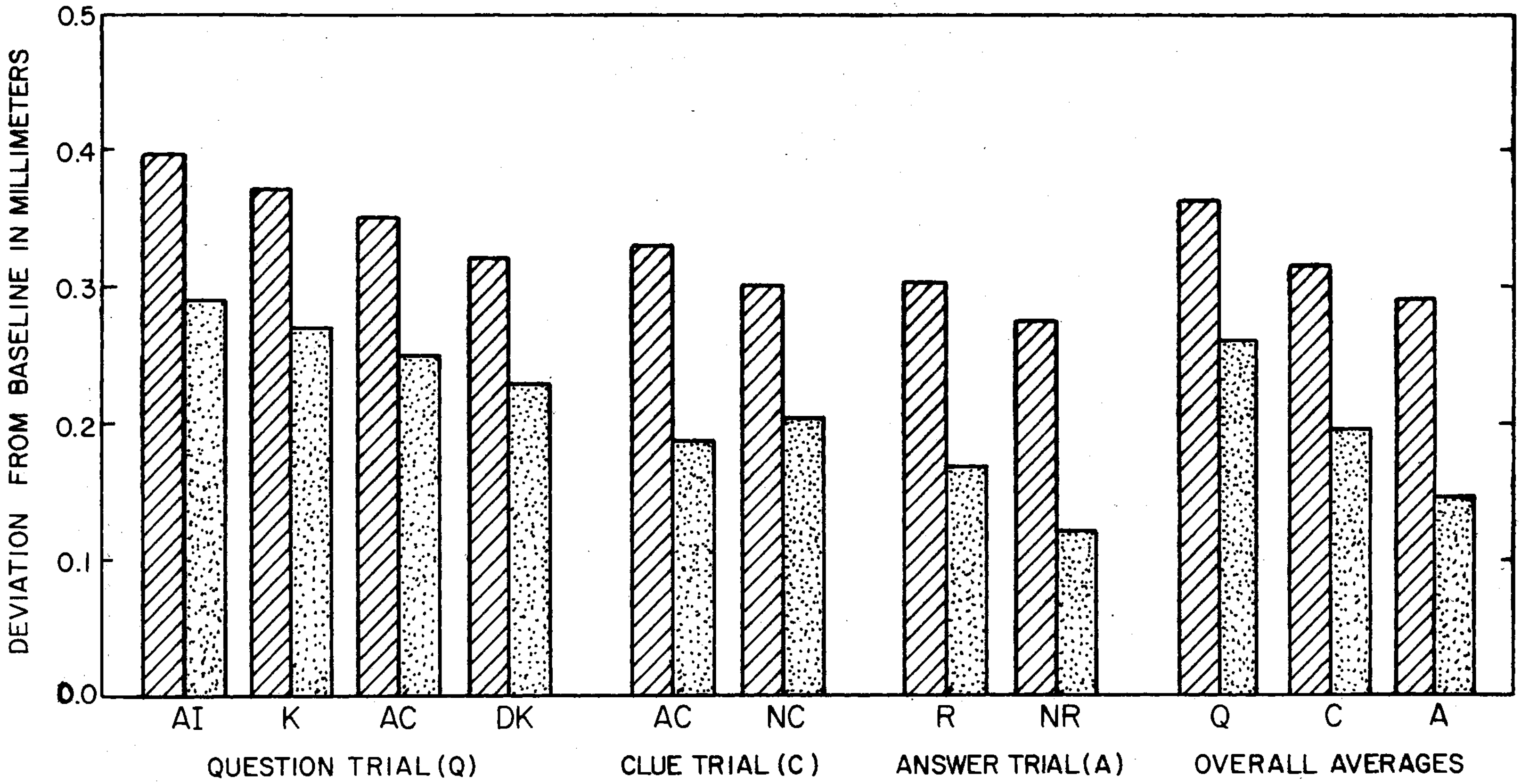


Figure 3. Average Peak Pupil Size and Average Processing Phase Pupil Change to Treatments

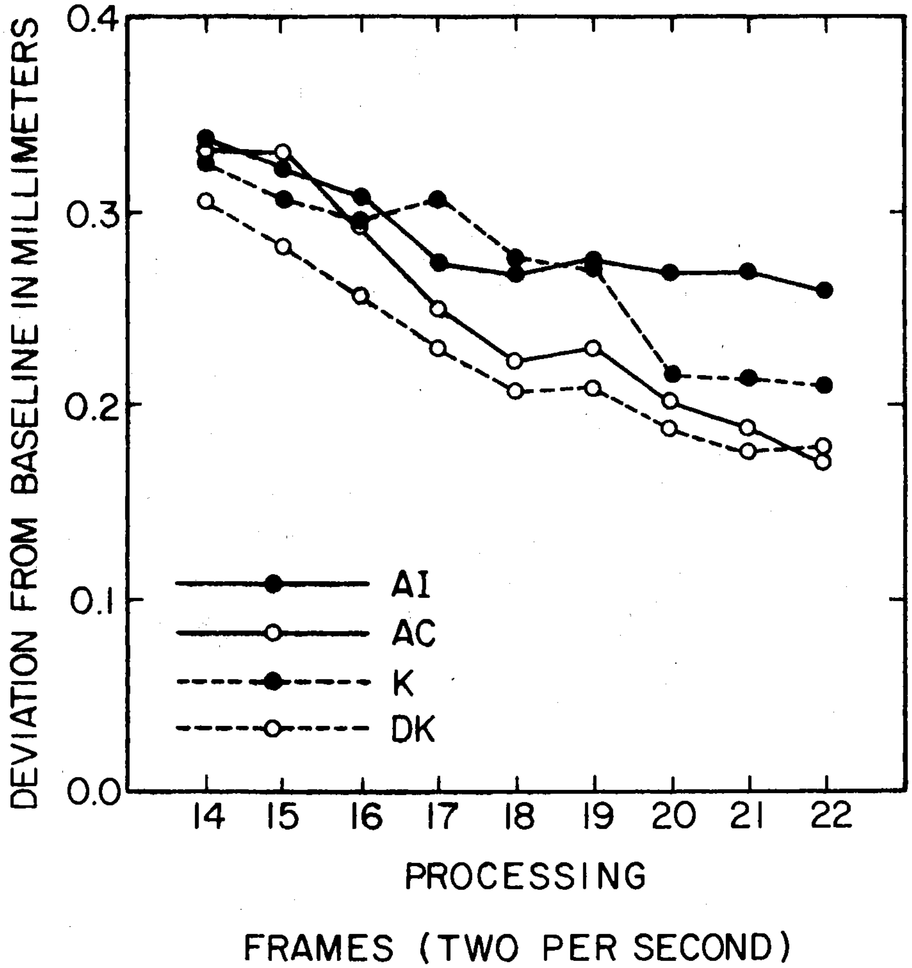


Figure 4. Processing Phase of the Four Outcomes to the Question Trial

TABLE IV

AOV OF PUPIL RESPONSE TO TREATMENTS DURING PROCESSING PHASE OF QUESTION TRIAL

Source	df	MS	F (df corrected)
Total	431		
Subjects (S)	11	0.74568	
Frames (F)	8	0.08898	13.95[b]
FS	88	0.00638	
Treatments (T)	3	0.07751	3.30
TS	33	0.02350	
FT	24	0.00403	1.84
FTS	264	0.00219	

TABLE V

AOV OF PEAK PUPIL RESPONSE TO TREATMENTS
OF QUESTION TRIAL

Source	df	MS	F (df corrected)
Total	47		
Subjects (S)	11	0.08894	
Treatments (T)	3	0.01254	5.31[e]
TS	33	0.00236	

TABLE VI

AOV OF LATENCY TO PEAK RESPONSE OF
QUESTION TRIAL TREATMENTS

Source	df	MS	F (df corrected)
Total	47		
Subjects (S)	11	6.88447	
Treatments (T)	3	4.90972	2.699
TS	33	1.81881	

latency to be longer than the DK latency. It might be noted that the rank order of these four means is the same as that for the peak means and overall means (Fig. 3).

Clue Trial

The time series for the AC and "No Closer" (NC) categories are shown in Figure 5 (It will be recalled that the AI and "Closer" ratings were excluded from analysis.). There is no difference between the curves during the Listening phase ($p > .20$) as well as the Processing phase ($p > .20$, Table VII). However, the difference at the first frame of the latter phase (# 12) is significant (.326 mm vs .279 mm; $t = 2.610$; $p < .05$); this difference parallels that found at the second frame of the Processing phase for the AC and DK categories of the Question trial. The differences may be reflective of retrieval vs non-retrieval of a potential response candidate. This point will be elaborated in the discussion section.

Both the peak-response measure ($t = 1.764$, $p < .2$) and the latency measure ($t = 0.584$) showed no differences between the two categories, as was the case for AC and DK of the Question trial (Although Fig. 5 suggests that frame 12 is the peak for both curves, not all the Ss' peaks occurred at this point, especially for the NC condition.).

Answer Trial

The data for the "Recognize" (R) and "Not Recognize" (NR) ratings are presented in Figure 6. An analysis of the curves during the Processing phase shows a significant difference between R and NR ($p < .05$, Table VIII). This result must be approached with caution, however, due

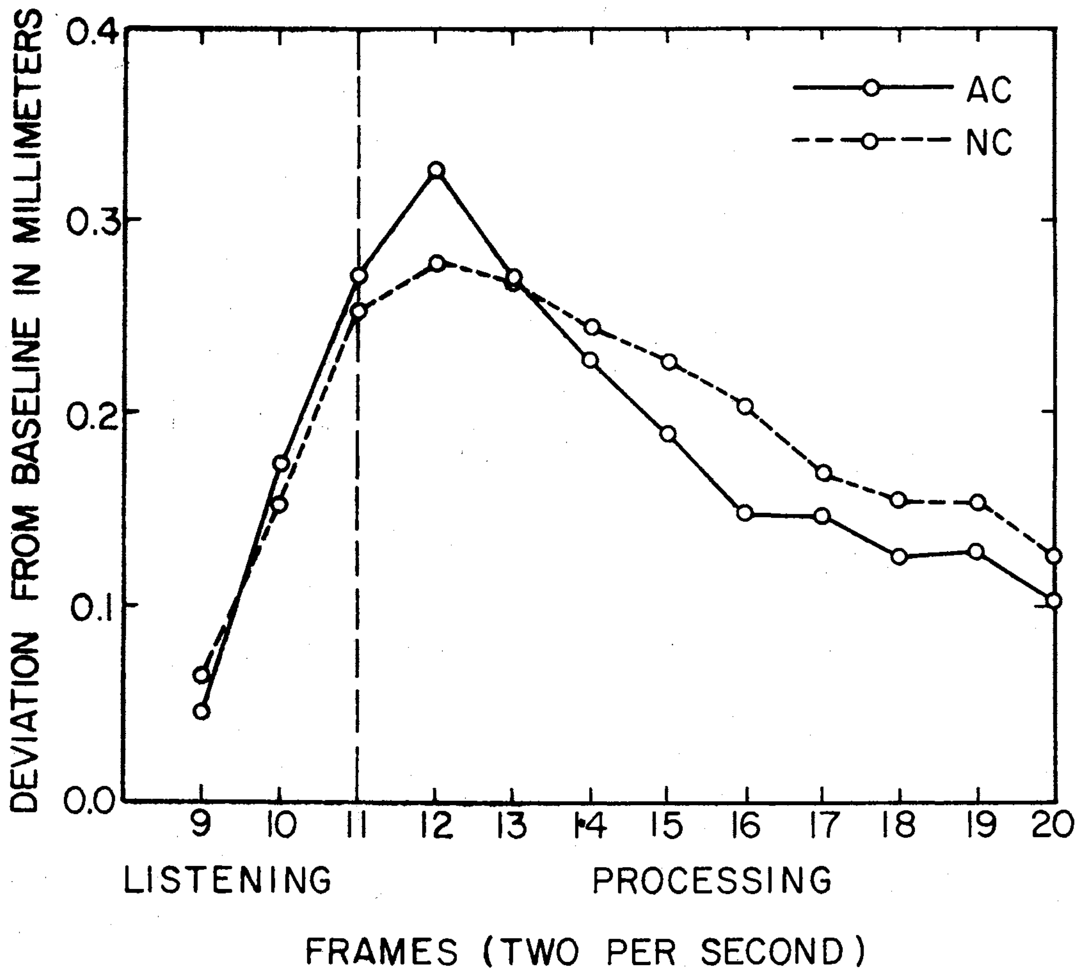


Figure 5. Pupil Response to Answer Correct (AC) and "No Closer" (NC) Outcomes of the Clue Trial

TABLE VII

AOV OF PUPIL RESPONSE DURING PROCESSING PHASE
TO AC AND NC OF CLUE TRIAL

Source	df	MS	F (df corrected)
Total	215		
Subjects (S)	11	0.28328	
Frames (F)	8	0.09722	35.35[a]
FS	88	0.00275	
Treatment (T)	1	0.01554	< 1
TS	11	0.02006	
FT	8	0.00492	2.20
FTS	88	0.00224	

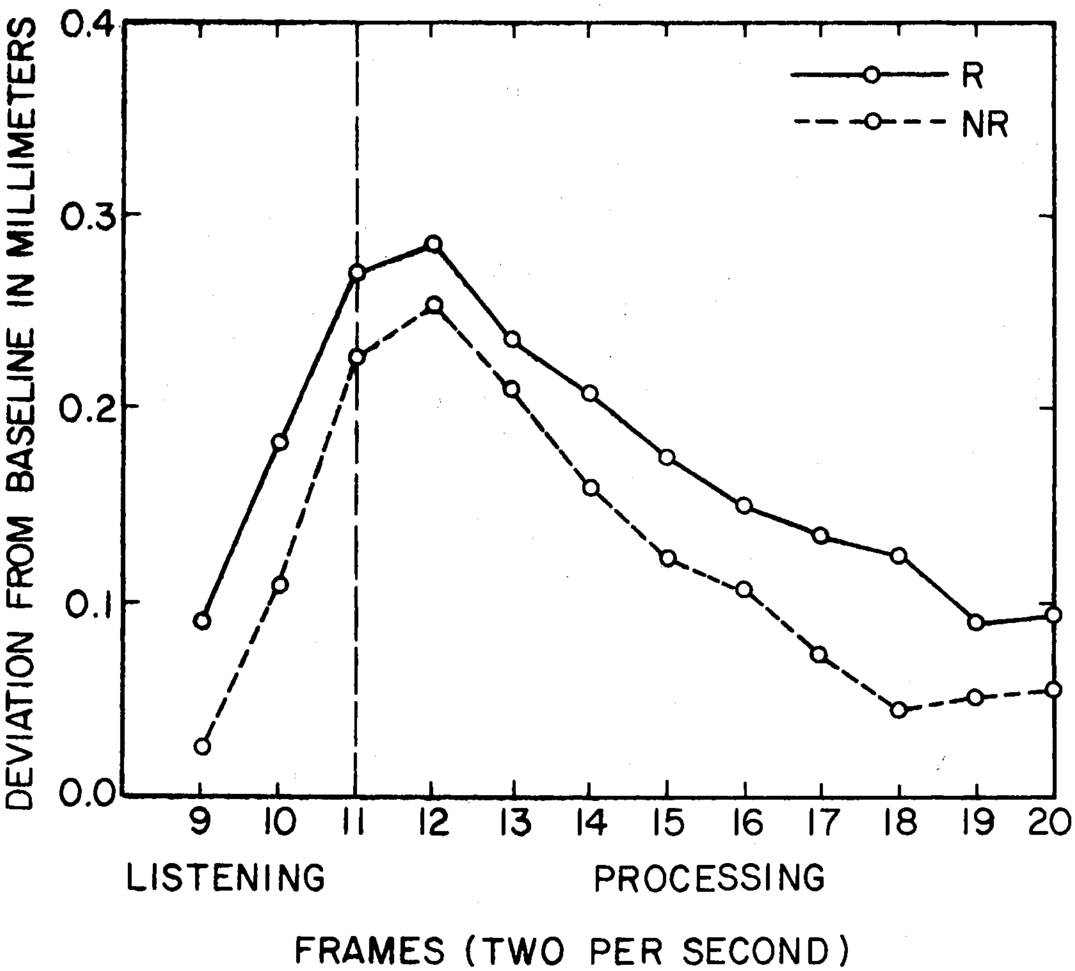


Figure 6. Pupil Response to "Recognize" (R) and "Not Recognize" (NR) Outcomes of the Answer Trial

TABLE VIII

AOV OF PUPIL SIZE DURING PROCESSING PHASE TO R AND NR OF ANSWER TRIAL

Source	df	MS	F (df corrected)
Total	215		
Subjects (S)	11	0.23963	
Frames (F)	8	0.11718	25.04[a]
FS	88	0.00468	
Rating (R)	1	0.11961	5.19[e]
RS	11	0.02302	
FR	8	0.00160	1.17
FRS	88	0.00137	

to a difference found during the Listening phase. The values are of different magnitude during this phase ($p < .025$, Table IX), and the result is puzzling because differences are not expected as the answer is being presented (The difference between R and NR begins with the first post-baseline frame (# 9, .093 mm vs .025 mm; $t = 3.310$, $p < .01$). All analyses of the Listening phases for the Question and Clue trials produced very small F-ratios, in line with the expectation that differences would not occur as information was being received.

A stronger argument could be made for the pupil's ability to distinguish the R and NR ratings if the curves at least ended the Listening phase at the same magnitude (A look at values for the other curves shows this to be the case; no differences at the end of the period were significant.). Since the Frames-by-Rating interaction was not significant in either phase, the pattern of the curves is essentially the same, and it is conceivable that the R curve would fall on top of the NR curve had it begun at the same magnitude. This possibility is strengthened by the fact that the average amount of change during the Listening phase (difference between frames 9 and 11) is similar for the two curves ($t = 1.707$, $p < .20$). Likewise, the average drop during the Processing phase (difference between the high and low point) showed no difference ($t = 1.041$, $p < .40$).

In spite of these negative points concerning the plausibility of the R-NR difference during the Processing phase, certain positive points exist in the data as well. The possibility of widely different baselines between R and NR was examined by calculating the average, absolute pupil size for each of the eight baseline frames. These untransformed values are shown in Figure 7, panel a (the first frame of the Listening phase is

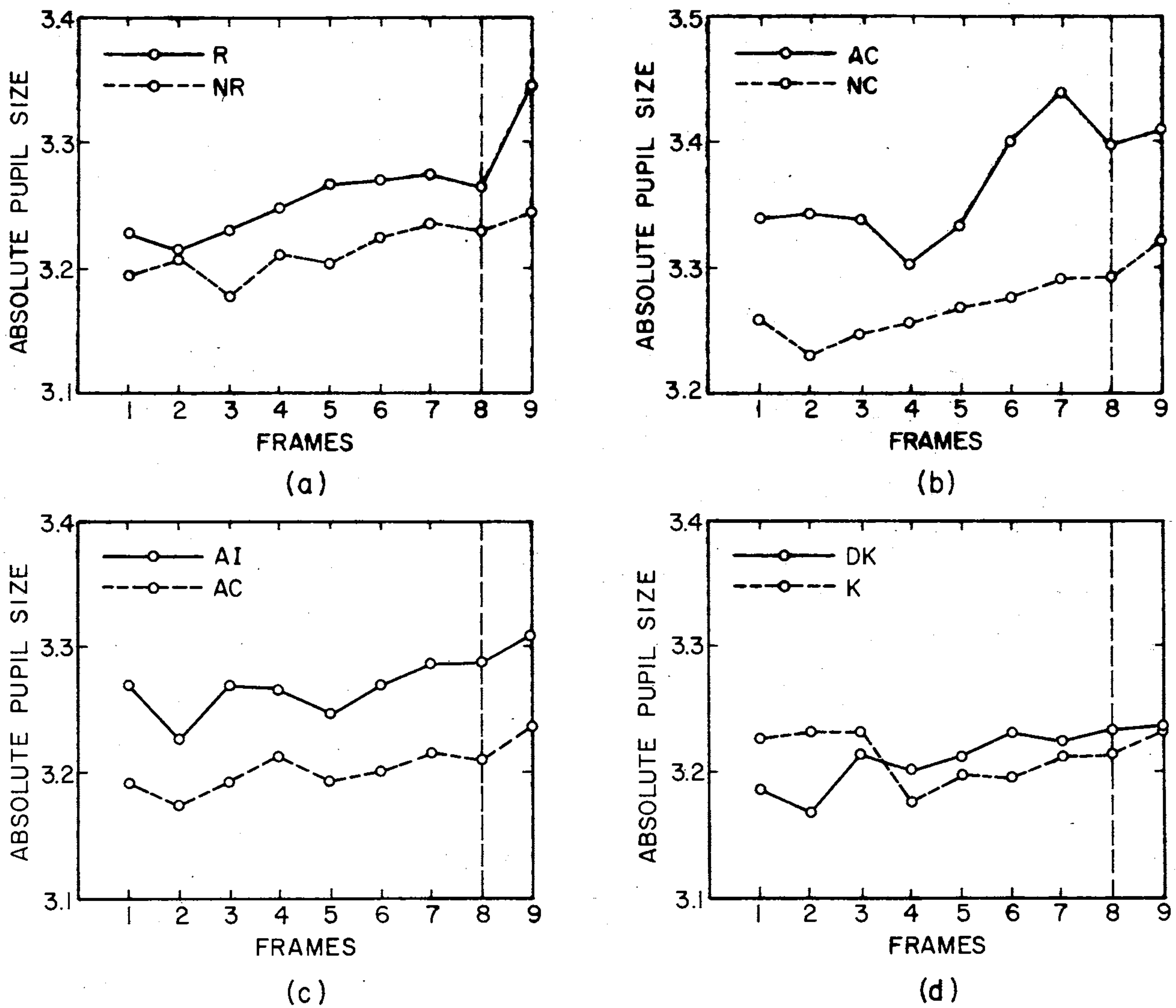


Figure 7. Average Absolute Pupil Size During Baseline Period for Answer Trial (a), Clue Trial (b), and AC vs AI (c) and K vs DK (d) of Question Trial

TABLE IX

AOV OF PUPIL SIZE DURING LISTENING PHASE TO
R AND NR OF ANSWER TRIAL

Source	df	MS	F (df corrected)
Total	71		
Subjects (S)	11	0.03892	
Frames (F)	2	0.21910	38.57[a]
FS	22	0.00568	
Rating (R)	1	0.06686	7.25[d]
RS	11	0.00922	
FR	2	0.00140	1.54
FRS		0.00091	

presented as well). Similar graphs for AC (Clue) vs NC, AC (Question) vs AI, and K vs DK are also presented for comparative purposes (panels b, c, & d, respectively). All the curves (with the possible exception of AC-Clue) exhibit smooth patterns over the four second baseline period, and appear to fluctuate within the limits of "pupil unrest" (any difference between baselines is not a factor here since a baseline was recorded on every trial).

The most striking feature of the graphs is the large increase in pupil size from frames 8 and 9 of the R curve. Although an increase exists in all the graphs at frame 9, that of R is the greatest. This jump could indicate an anticipation effect for an answer one felt he would recognize, even though he could not previously supply the answer. A S might therefore be more receptive to the stimulus, and show higher dilation throughout the Listening phase. To check on this anticipation notion, the R trials of 8 of the Ss were divided into two sets: one set contained those R trials whose corresponding rating on the Question trial was K (K-R), and one set whose previous rating was DK (DK-R; the sets of 4 Ss' data were not used because they did not fit the criterion of at least 2 trials in each set). If Ss were anticipating some of the answers, it would seem that they would be more apt to have a feeling for those which they rated they knew, but could not quite produce. Thus, it was expected that the K-R curve would be higher than the DK-R curve during the Listening phase. Although the curves do show a difference in the hypothesized direction during this phase, an AOV indicated that the difference was not significant ($p < .20$, Table X; there was also no difference in pupil size during the Processing phase, $p < .20$). Hence, the anticipation hypothesis is only suggestive, and is not statistically

TABLE X

AOV OF PUPIL SIZE DURING LISTENING PHASE TO
K-R AND DK-R OF ANSWER TRIAL

Source	df	MS	F (df corrected)
Total	47		
Subjects (S)	7	0.04829	
Frames (F)	2	0.17914	27.90[b]
FS	14	0.00642	
Set (T)	1	0.02646	2.16
TS	7	0.01226	
FT	2	0.00260	1.18
FTS	14	0.00220	

substantiated.

Although the R and NR curves did differ during the Listening phase, the NR curve "caught up" to the R curve at the beginning of the Processing phase (Fig. 6). Neither frames 12 ($t = 1.711$, $p < .20$), 13 ($t = 1.228$, $p < .30$), or 14 ($t = 1.671$, $p < .20$) differed significantly. The separation began two seconds after reception of the answer (# 15; $t = 2.678$, $p < .05$). The pupil response to R ratings, then, maintained a larger size during the latter portion of the Processing phase, rather than dropping at an identical rate with the NR curve.

This result and the possible anticipation effect weigh as positive factors in judging the reliability of the Processing-phase difference between R and NR.

Neither the peak-dilation measure ($t = 1.602$, $p < .20$) nor the latency measure ($t = 0.499$) differentiated the ratings during the Processing phase. However, these two results are not necessarily in conflict with the overall difference between R and NR, because as just explained, the separation does not begin with the onset of the Processing phase.

A possible confound of the DK trials was discussed earlier. A similar confound could exist in NR ratings. All the answers to the questions were intended to be ones with which the S was familiar. However, the pupil might respond differently to an answer one has heard before, but does not connect with the question, as compared to an answer he encounters for the first time. A list of all the answers was presented to each S after the experiment, and he was instructed to check any novel answers. Since these checks accounted for 33% of the NR trials for the 12 Ss, the "novel" NR trials were separated from the other NR trials, and an AOV was performed on the two time series. The difference was negligible ($p > .20$), and hence it was concluded that the NR trials were not differentially affected by "novel" vs "non-novel" types of NR trials (all NR trials were used in the analyses previously discussed).

Inter-Trial Comparisons

AC-Question vs AC-Clue vs R

In order to compare the amount of dilation to positive outcomes of the three trial-types, the Processing phase of the two AC conditions and the R conditions (Fig. 8, panel a) were analyzed together (Table XI). The resulting F-ratio indicated the possibility of differences ($.05 < p < .10$), and Newman-Keuls procedures on the means (AC-Question: .247 mm; AC-Clue: .187 mm; R: .168 mm) indicated that the AC-Question curve was of greater magnitude than both the AC-Clue and R curves ($p = .05$). The last Listening-period frames were compared by *t* tests, and the three treatments ended that period at essentially the same magnitude (no tests were significant). Hence, the major differences in pupil size occurred after the intake of the three different types of information.

DK vs NC vs NR

These three curves, reflecting a negative outcome for the three trial types (Fig. 8, panel b), were likewise analyzed together (Table XII), and a significant difference was found among the treatments ($p < .01$). A Newman-Keuls test ($p = .05$) showed both the DK and NC means (.227 mm; .203 mm) to be greater than the NR mean (.121 mm), but not greater than each other. The three treatments also ended the Listening phase at the same level (no *t* tests were significant).

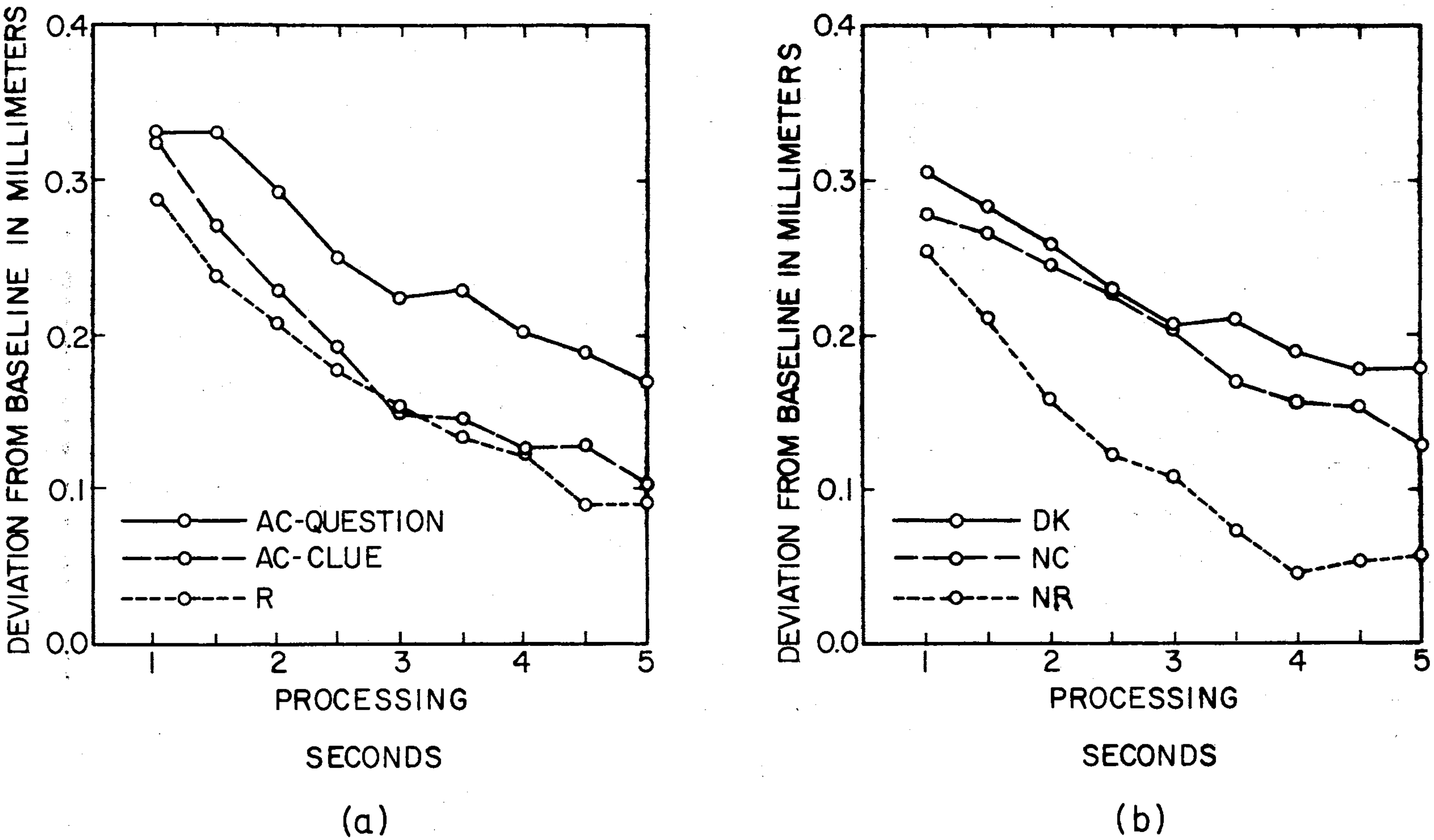


Figure 8. Pupil Size During Processing Phase to Comparable "Positive" Outcomes (a) and "Negative" Outcomes (b) of the Three Trial Types

TABLE XI

AOV OF PUPIL SIZE DURING PROCESSING PHASE TO AC-QUESTION, AC-CLUE, AND R TRIALS

Source	df	MS	F (df corrected)
Total	323		
Subjects (S)	11	0.36850	
Frames (F)	8	0.15979	53.09[a]
FS	88	0.00301	
Treatments (T)	2	0.18340	4.36
TS	22	0.04210	
FT	16	0.00190	< 1
FTS	176	0.00231	

TABLE XII

AOV OF PUPIL RESPONSE DURING PROCESSING PHASE TO
DK, NC, AND NR TRIALS

Source	df	MS	F (df corrected)
Total	323		
Subjects (S)	11	0.39315	
Frames (F)	8	0.11786	32.03[a]
FS	88	0.00368	
Treatments (T)	2	0.33337	12.05[c]
TS	22	0.02766	
FT	16	0.00402	2.05
FTS	176	0.00196	

Question vs Clue vs Answer Trial

In order to obtain an overall picture of the degree of mental effort during the Processing phase for each of the three tasks (Question: retrieval; Clue: problem solving; Answer: decision), the overall means of each trial-type/S were used as the units for analysis. The averages over Ss are shown in Figure 3 (dotted bars) and as the graph suggests, there is a significant difference in overall effort to each task ($p < .005$; Table XIII). Newman-Keuls comparisons ($p = .05$) indicated that the Question trial is of greater magnitude than both the Clue and Answer trials, and that the Clue trial is greater than the Answer trial.

A similar AOV on the average peak response of each trial type showed that this measure also differentiated the tasks ($p < .01$, Table XIV; Fig. 3, striped bars): the Question trial peaked at a greater magnitude than both the Clue and Answer trials (Newman-Keuls, $p = .05$).

The latency measure was not analyzed for this inter-trial comparison because the time of the peak was related to the length of the Listening phase. On the average, it occurred approximately two frames into the Processing phase of the Question trial (5 frame Listening phase), and on approximately the first frame of the Processing phase of the Clue and Answer trials (3 frame Listening phase).

In order to compare the amount of dilation of the three trial types during the Listening period, the average peak for each type/S was calculated (the mean of the Listening phase would not serve as a suitable statistic because the values of the Clue and Answer trials rise rapidly from frame-to-frame, whereas the major increases during the longer Question trial do not occur until the latter part of the period. An AOV showed very little difference among the types ($p > .20$, Table XV),

TABLE XIII

AOV OF PUPIL SIZE DURING PROCESSING PHASE
TO QUESTION, CLUE, AND ANSWER TRIALS

Source	df	MS	F (df corrected)
Total	35		
Subjects (S)	11	0.04429	
Task (T)	2	0.03874	14.20[b]
TS	22	0.00273	

TABLE XIV

AOV OF PEAK PUPIL SIZE TO QUESTION, CLUE, AND ANSWER TRIALS

Source	df	MS	F (df corrected)
Total	35		
Subjects (S)	11	0.05466	
Task (T)	2	0.01625	9.91[c]
TS	22	0.00164	

TABLE XV

AOV OF PEAK PUPIL SIZE DURING LISTENING PHASE TO QUESTION, CLUE, AND ANSWER TRIALS

Source	df	MS	F (df corrected)
Total	35		
Subjects (S)	11	0.03824	
Task (T)	2	0.00238	1.18
TS	22	0.00202	

indicating that the high point of dilation was similar for the intake of the three types of information.

A Check on Law of Initial Values

Because both the average basal sizes and average magnitudes of responding varied to some extent (basal size: 2.86 mm to 4.25 mm; change in Processing phase: .065 mm to .461 mm), tests were performed on prestimulus values. The Law of Initial Values (LIV), discussed earlier in the statistical section of Chapter II, deals with changes from a baseline level as the response measure (Wilder, 1967, pp. 25-27), and one way it has been statistically defined is the product moment correlation between the prestimulus levels and the change scores (e.g., Benjamin, 1963, p. 558; Oken & Heath, 1963, pp. 4-5). For the purpose of studying the present data, the average basal size of each S/condition was paired with his average peak score for that condition; the peak dilation (range of average over treatments = .142 mm to .618 mm) was used as the change score because such scores are the most sensitive measure; that is, how high a level of dilation can the pupillary system attain before opposing mechanisms attempt to counteract the change.

The eight correlation coefficients are presented in Table XVI. All were significant, indicating that the peak magnitude of response was somewhat a function of the prestimulus level. The smaller peak changes were associated with the lower basal sizes, whereas the greater changes were associated with the larger basal sizes. Also presented in the table are the following statistics: 1) r^2, which represents the strength of the linear relationship; 2) $1-r^2$, which represents the proportion of variance still remaining after the basal sizes are accounted for; and 3)

TABLE XVI

PRODUCT-MOMENT CORRELATIONS BETWEEN BASAL AND PEAK PUPIL SIZE

Statistic	Question				Clue		Answer	
	AC	AI	K	DK	AC	NC	R	NR
r	.68[d]	.79[c]	.83[a]	.78[c]	.86[a]	.84[a]	.84[a]	.87[a]
r^2	.46	.62	.69	.61	.74	.71	.71	.76
$1-r^2$	.54	.38	.31	.39	.26	.29	.29	.24
b	.26	.40	.37	.27	.24	.31	.30	.34

the slope, or regression coefficient, which expresses how much of an increase in change score would be expected to accompany each increment in basal size (these values are of course different from zero since the r's were significant).

The LIV was explained in Chapter II as indicating that larger prestimulus levels were associated with smaller changes. This fact implies that normally the correlation between baseline and change is negative. In addition, however, the Law states that a relative lack of responding occurs at extreme levels. In the present study, the lower prestimulus pupil sizes were somewhat near the physiologically lower bound of 2 mm (e.g., 5 of the 12 Ss' overall average baselines were under 3.1 mm). The larger sizes were distant from this lower bound, and also well below the upper bound of 8 mm (e.g., 4.3 mm), and hence were "freer" to increase without opposing forces coming into play. Hence, the lower levels were associated with less change, and the resulting correlations were positive.

It may be assumed that basal sizes near the upper bound would show as little change as those near the lower bound; in fact, Kahneman et al. (1967, Fig. 1a, b) show pupil response curves from an experiment in which illumination of the eye was purposely set at a low level, thereby causing larger initial basal sizes (about 6.75 mm). The pupil responded in a "sluggish" manner to a task which produced larger dilations at a higher illumination level (and therefore lower basal size -- about 4.0 mm).

One process which has been invoked to explain the LIV is that of homeostasis (e.g., Sternbach, 1966). The reasoning as applied to pupil size is that a response away from an attained equilibrium activates a mechanism to return the organ to its previous state. Apparently, this mechanism is more sensitive at extreme levels (the closer to the lower limit, say, the greater the constrictive effect works to offset a dilation and return the pupil to its initial state).

The reasons why basal sizes may differ between subjects are as follows: "(1) the individuals are not in the same behavioral state, and (2) they are in the same behavioral state, but their physiological characteristics are dissimilar" (Heath & Oken, 1965, p. 467). The second reason would seem to explain why the levels did differ in spite of approximately equal illumination levels for all Ss. The main concern, however, is the interpretation of the results which have been presented in this chapter, because "If the magnitude of autonomic response to a stimulus is dependent upon the initial, or prestimulus level, how can one fairly compare individuals' or groups' response to stimulation if they have differeing initial levels?" (Benjamin, 1963, p. 557). It will be shown below that although there were differences among the S's in magnitude of responding, the Ss were similar in the pattern of their responses to the various treatments.

The extent to which the rank ordering of response to the eight treatments tended to be similar was determined by calculating Kendall's W statistic (coefficient of concordance). The means of the Processing phase for each condition/S were used for the ranking procedures; the object was to determine whether the apparent dependence of magnitude of responding on initial level affected the patterns of sustained responding (that is, over the length of the Processing phase).

The W value was .49 (on a 0 to 1 scale), indicating a degree of similarity of rank orderings among the Ss. A significance test of this relationship (Hays, 1963), p. 658) showed a significant ordering of ranks ($p < .001$), and therefore it was concluded that Kendall's W was not equal to zero. Thus, in spite of different magnitudes of responding, the pattern of response by the Ss was similar to the ordering of the treatments as shown in Figure 3.

Another statistical method to show that Ss preserved similar relative positions across treatments is the intraclass correlation, which tests specific variance components. In the case of the present data, the relationship is tested by $R = [MS_{subs.} - MS_{subs.\ x\ trts.}]/[MS_{subs.} + (k-1)\ MS_{subs.\ x\ trts.}]$, where k = the number of categories (Wilson, 1967, p. 136; see also Myers, 1966, pp. 294-299). This formula tests the strength of the trend for observations in a given category to be more like each other than observations in different categories. In order to test the degree of this relationship, the means of the eight Processing phases for each S were included in an AOV (Table XVII), and the error components were applied to the formula. R was equal to .80, and implies that $MS_{subs.}$ (which reflects the difference in magnitude of responding between Ss) accounted for relatively more variance than did $MS_{subs.\ x\ trts.}$ (which measures the failure of Ss to act alike to the various treatments).

The ordering of the Ss' responses to the Question, Clue, and Answer trials as a whole were also similar. Kendall's W was .60, indicating a fair agreement in rankings among the Ss, and a significance test verified this relationship ($p < .001$). Furthermore, the intraclass correlation was .83 (components taken from Table XIII), providing further evidence that the Ss' individual responses were in accord with the ordering shown in Figure 3.

Many statistical techniques of data adjustment to correct for the initial value - response relationship have been proposed. These

TABLE XVII

AOV OF PUPIL SIZE DURING PROCESSING PHASE
TO THE EIGHT TREATMENTS

Source	df	MS	F (df corrected)
Total	95		
Subjects (S)	11	0.12698	
Treatments (T)	7	0.03688	9.81[c]
TS	77	0.00376	

techniques are designed to account for the degree to which the initial value has affected the results. However, no single approach has yet been decided upon (Lacey, 1956, was the first to suggest data transformations; discussion and new approaches to his suggestions may be found in Oken & Heath, 1963; Benjamin, 1963; Steinschneider & Lipton, 1965; Heath & Oken, 1965; Lykken, 1972; Sternbach, 1966; Van Egeren, Headrick & Hein, 1972; and Wilder, 1967). Because the initial-value effect did not seem to alter the pattern of response to treatments in the present experiment, and because all tests were within-Ss, data transformation was not deemed appropriate. Hence, the original magnitude, direction and time course of the responses are preserved.

Questionnaire Data

The percentages of trials falling in a given treatment are shown in Table I. Estimations of Ss' accuracy in monitoring their memories were made by comparing the eventual outcomes of K and DK ratings (Question trial). A mean of thirty-four percent of the K trials were answered correctly on the Clue trial, whereas only 15% DK trials were answered correctly on this trial. A t-test on the Ss' proportions for these two comparisons showed a significant difference ($t = 4.479$, $p < .001$). Of the remaining K trials, 69% were then rated as "Recognize" on the Answer trial, compared to 40% for the remaining DK trials ($t = 3.474$, $p < .01$). This latter comparison is equivalent to the tests Hart (1965b, 1966, 1967a, b) made of memory-monitoring accuracy (previously discussed in Ch. I). He compared positive feeling-of-knowing items correctly recognized with negative FOK items correctly recognized, and on the basis of significant differences between the two percentages, concluded that his Ss were able to monitor accurately their memory states. The same interpretation would seem to follow in the present experiment -- more K trials were answered correctly on the Clue trial, and more of the remaining trials were rated as R on the Answer trial than the DK counterpart. Although the tasks in this and Hart's studies are not identical (Ss chose from a multi-item recognition test), the logic behind the comparisons is similar, and it appears that the Ss were able to evaluate the availability of information.

The data can be further divided to compare the Ss' accuracy of each type of rating. A comparison between the K's which were recognized (69%) versus those which were not recognized (31%) tests the accuracy in rating what information is in storage (Hart, 1965b). This difference is significant ($t = 2.371$, $p < .05$). The DK's which were not recognized (60%) versus those which were recognized (40%) tests whether Ss' were able to rate accurately what is not in storage. However, this difference is only marginally significant ($t = 1.979$, $.05 < p < .10$), indicating that Ss sometimes realized the connection between the answer and a question which they rated as a DK. It would seem, though, that their initial DK ratings were accurate indicators of their knowledge about the question at the time, because the pupil responses during the Processing phase significantly differed from those associated with K ratings.

CHAPTER V

DISCUSSION

The present experiment was a study of how people recall facts. No attempt was made to control what the Ss learned; rather, the study was designed to focus on the retrieval component of memory. The nature of the task -- responding to general information questions -- is more akin to the daily use of our memories than, say, a task in which Ss are required to produce a list of recently memorized words. The emphasis of the pupillary response interpretations will therefore be geared to long-term memory retrieval theory.

One of the remarkable capabilities of the human memory is the ease with which we determine whether we have knowledge about a given query. The search process usually recovers fairly quickly the appropriate response to a question we know. Conversely, we can rapidly judge when the contents of our memories do not include certain information. Just as we quickly reject "mantiness" as being a word (Norman, 1969, p. 162), so too we rapidly "know that we do no know" the answer to a particular question. Complete lack of knowledge about a question, then, should result in few searches through the memory store. The pupillary responses to questions which resulted in a correct answer or a "Do Not Know" rating are in line with these notions. The curves for the most part are similar, the one exception being the difference near the beginning of the Processing phase. This point could represent the direct retrieval of the answer. Because one had certainty in his answer (prompt retrieval of an item may often be taken as a sign of confidence in it), search processes were terminated, and pupil size thereafter corresponded with the DK curve: certainty of lack of knowledge about a query appeared to result in an early decision to terminate search. If a rapid search through memory found no related information at all, further attempts would not be useful. Part of the efficiency of the memory system depends on the accuracy of our "lack of knowledge" judgments about an item. Such efficiency is verified by the feeling-of-knowing research discussed earlier.

A similar pattern of response occurred to the AC and NC counterparts of the Clue trial. The sharp drop of AC after frame 12 and the steady decline of AC after frame 15 of the Question trial are in line with post-report pupillary drops during the unloading phase of a short-term memory task (Kahneman & Beatty, 1966) as well as the post-solution declines to single-answer anagram task (Bradshaw, 1967) or multiplication task (Hess & Polt, 1964); also, Hess (1972, p. 523) has reported little pupil change to mental problems which are easily solved or which are judged too difficult to solve. The difference with the present tasks is that no overt response is made until the end of the Processing period. The declines may therefore be reflective of a drop in processing (because of answer achievement for AC and search termination for DK).

Uncertainty about a candidate for recall leads to decisional processes on the item. The difference between AI and AC during the last two seconds of the Processing phase would seem to verify this notion. The larger pupil size of AI may reflect an editing process on one implicitly retrieved response, or a decision between two potential responses. The S must decide whether the recovered information is appropriate or consistent within the context of the presented information. In either event, the pupillary response was different from the cases in which 1) direct retrieval of the desired information was achieved (and the S had confidence in it: AC), and 2) search termination probably occurred fairly quickly (DK).

Furthermore, a decision on a retrieved item (AI) resulted in a more sustained pupil size than a decision on attributes, or a feeling-of-knowing about an item (the difference between AI and K during the last one and a half seconds of the Processing phase). Also, knowledge of availability of an item (K) resulted in a larger pupil size than knowledge of unavailability (DK). Knowledge that an item is in storage may have caused a more prolonged search effort.

The nature of the three tasks caused differences in the amount of processing. The information contained in a question elicited search for a target which fit a particular category (e.g., an author, or a city). The Clue trial called for a more specific search -- a member of a search set which fit within the structural boundaries of the first and last letters. An Answer trial required a decisional process in the form of an association or nonassociation between question and answer. The overall pupil size during the Processing period showed a respective decrease as a function of the type of task. This difference existed in spite of the fact that the high point of pupil size did not vary during the Listening phase, where most of the dilation per se took place. The trend of the response during Processing was a return toward baseline, but the degree of this trend varied as a function of task requirements (since an overt response was given at the end of each trial, the trends were not a function of whether or not a response was required). Overall, pupil size remained at a higher level in the Answer trial than the other two, and higher during the Clue trial than the Answer trial.

It does not seem that an emotional or arousal component was a confound in the data. The most logical treatment in which it might appear is DK. There could be anxiety over experimenter evaluation when it became obvious to the S that he did not know the answer to a given question. The emotional component would presumably manifest itself in an increased pupil size (e.g., see Simpson & Molloy, 1971). However, because the overall mean of DK is the lowest of the Question treatments, an interpretation of the pupil response in terms of an emotional hypothesis is not justified. Rather, pupil size reflected the momentary processing load on the S: the highest average pupil size during Processing occurred to the AI treatment, whereas the lowest occurred to the NR condition; AI is concerned with the judging of potential response correctness, while NR represents more of an automatic, passive decision. Preparation for an overt response did not appear to have an increased effect on pupil size either, because all curves show a downward trend.

Thus, an explanation of the data in terms of the degree of mental processing seems more reasonable, as was the case in previously cited pupillary studies which showed the responding to be more in line with processing requirements than emotional considerations (see last part of "What Does the Pupillary Response Measure", Ch. II).

Given this interpretation, the pupillary data appears to substantiate certain aspects of current retrieval theories. The retrieval process has been divided by some researchers into a series of stages (Norman, 1970, 1972) or mechanisms (Shiffrin & Atkinson, 1969). A query is first preprocessed to determine whether any relationships between the items exist in memory. If so, the search process examines specific storage locations. On some occasions, information is directly recovered, and given an emit decision (AC). On others, the response-generation process judges the correctness of the retrieved item or items (AI). In some cases, search is able to select for examination only attributes of the desired information, and the post-processing component must decide how much relevant information has been retrieved (K). Should no relationships of the question be found in memory, search is unlikely to continue very long before being terminated (DK).

When the search process becomes essentially a problem solving task (Clue trial), the achievement of the correct answer does not result in as much processing as the achievement of the answer to the question itself (Fig. 8a). Perhaps since the constraints are greater on the Clue trial, "knowing-that-one-knows" feedback is greater. Or, because the clues were of aid on an all-or-none basis (relatively few AI and C categories), the initiation of any search was very specific and resulted in direct retrieval of the target. Search termination due to unsuccessful recovery of relevant information to the question or clues shows the same level of effort (Fig. 8b).

The recognition of an answer as being connected to the question resulted in as much processing as achieving the answer on the Clue trial (but not as much effort involved in direct retrieval to the question; Fig. 8a). Not realizing a connection resulted in the least processing of all (NR; Fig. 8b).

To the author's knowledge, this study is only the second pupillary experiment concerned with retrieval of information from long-term memory. In 1966, Beatty and Kahneman compared the effort involved in retrieving well-known telephone numbers to that involved in a short-term memory task. The majority of pupillary studies have used the short-term memory paradigm, and pupil size is monitored while material is presented and/or output. The present study made no attempt to control the Ss' memory storage. The monitoring of pupillary responses was primarily concerned with the retrieval aspects of memory. Pupil size appears to be a sensitive index of the amount of processing involved in the covert process of retrieval, because differences in dilation existed among the four Question trial outcomes. Also, the overall dilations to the Question, Clue, and Answer trials indicated processing differences to these three tasks.

CHAPTER VI

SUMMARY

The purpose of the study was to investigate the pupillary response to retrieval of factual information stored in long-term memory. Twelve subjects were each auditorially presented a series of general knowledge questions (e.g., "In what country is the Taj Mahal?"). If the S could not produce an answer, he rated whether he felt he did nevertheless know the answer, or did not know it. Questions not answered (correctly) were followed by a Clue trial in which the first and last letters of the target were presented. S again attempted recall, but in the absence of a response gave a binary rating of closeness to the answer ("Closer"/"Not Closer"). Failure to produce the answer then resulted in an Answer trial; Ss were to rate whether they did or did not recognize the connection between the question and answer (all presented answers were correct).

The major dependent variable was change in pupil size. Filmed records of the eye were taken during all trials, each of which included a baseline, Listening, and Processing period (Ss' overt responses were given at the end of the latter period).

The major findings were as follows: Pupil size was higher during the last 2 seconds of the 4 1/2 second Processing phase for questions answered incorrectly (AI) than for those answered correctly (AC). The difference was attributed to the confidence one had in his potential answer. The pupillary response to those questions give a "Know" (K) rating showed a larger size than to those questions rated "Do Not Know" (DK); the pupil size of AI trials was also larger than DK trials, and it was concluded that pupil size was separating those instances in which judgments about retrieved items (AI) or attributes (K) caused more processing than the decision to terminate search (DK).

The pattern of responding to AC and "No Closer" (NC) ratings of the Clue trial was similar, except for a larger AC pupil size at the beginning of the trial (AI and "Closer" categories rarely occurred, and were excluded from analysis). These two results also characterized the AC and DK outcomes of the Question trial, and it was reasoned that retrieval of a correct answer was direct and resulted in suspension of processing, such that pupil size was thereafter similar to that of the trials in which search was terminated because no recovery of information was forthcoming (DK and NC).

Pupil size to answers given a "Recognize" rating was larger than to those given "Not Recognize" ratings. The pupillary response to this latter category was the lowest of all the categories, and it seems that very little processing was involved when an answer meant nothing to the S in regard to the question.

The K and DK ratings appeared to be fairly accurate indicators of the Ss' knowledge, because more K than DK rated questions were then 1) answered correctly on the Clue trial, and 2) given a "Recognize" rating on the Answer trial.

Overall, pupil size showed a significant progressive decrease to the Question, Clue, and Answer trials. The differences were attributed to the nature of the tasks. Because pupil size in this study was able to differentiate task demands as well as within-task outcomes, it was concluded that the pupillary response was a sensitive index of processing efforts. The technique of pupillometry offered a second-by-second monitoring of the covert retrieval process, and the findings fit within the framework of retrieval theory.

A SELECTED BIBLIOGRAPHY

Adler, F. H. Physiology of the eye. (3^{rd} Ed.) St. Louis: Mosby, 1959, pp. 167-210.

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

Adams, J. A. Response feedback and learning. Psychological Bulletin, 1968, 70, 486-504.

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

Ax, A. F. Goals and methods of psychophysiology. Psychophysiology, 1964, 1, 8-25.

Beatty, J., & Kahneman, D. Pupillary changes in two memory tasks. Psychonomic Science, 1966, 5, 371-372.

Benjamin, L. S. Statistical treatment of the law of initial values (LIV) in autonomic research: A review and recommendation. Psychosomatic Medicine, 1963, 25, 556-566.

Bernbach, H. A. Decision processes in memory. Psychological Review, 1967, 74, 462-480.

Bernick, N., & Oberlander, M. Effect of verbalization and two different modes of experiencing on pupil size. Perception & Psychophysics, 1968, 3, 327-330.

Bradshaw, J. Pupil size as a measure of arousal during information processing. Nature, 1967, 216, 515-516.

Bradshaw, J. Background light intensity and the pupillary response in a reaction time task. Psychonomic Science, 1969, 14, 271-272.

Bradshaw, J. Pupil size and drug state in a reaction time task. Psychonomic Science, 1970, 18, 112-113.

Brown, R., & McNeill, D. The "tip of the tongue" phenomenon. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 325-337.

Clark, W. R. A pupillographic study of short-term memory search. Unpublished doctoral dissertation, Oklahoma State University, 1970.

Clark, W. R., & Johnson, D. A. Effects of instructional set on pupillary responses during a short-term memory task. Journal of Experimental Psychology, 1970, 85, 315-317.

Clarke, F. R., Birdsall, T. G., & Tanner, W. P., Jr. Two types of ROC curves and definition of parameters. Journal of the Acoustical Society of America, 1959, 31, 629-630.

Colman, F. D., & Paivio, A. Pupillary response and galvanic skin response during an imagery task. Psychonomic Science, 1969, 16, 296-297.

Colman, F., & Paivio, A. Pupillary dilation and mediation processes during paired-associate learning. Canadian Journal of Psychology, 1970, 24, 261-270.

Danford, M. B., Hughes, H. M., & McNee, R. C. On the analysis of repeated-measurements experiments. Biometrics, 1960, 16, 547-565.

DaPolito, F., Guttenplan, H., & Steinitz, H. Accuracy of subjective judgments of information in long-term memory. Psychonomic Science, 1968, 13, 227-228.

Davson, H. The physiology of the eye. (2nd Ed.) Boston: Little, Brown and Co., 1963.

Fillenbaum, S. Psycholinguistics. Annual Review of Psychology, 1971, 22, 251-308.

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. Developmental changes in memorization processes. Cognitive Psychology, 1970, 1, 324-340.

Fredericks, R. S., & Groves, M. Pupil change and stimulus pleasantness. Proceedings of the 79th Annual Convention of the American Psychological Association, 1971, 6, 371-372.

Freedman, J. L., & Landauer, T. K. Retrieval of long-term memory: "Tip-of-the-tongue" phenomenon. Psychonomic Science, 1966, 4, 309-310.

Freud, S. Psychopathology of everyday life. New York: Macmillan, 1955.

Garner, W. R., Hake, H. W., & Eriksen, C. W. Operationism and the concept of perception. Psychological Review, 1956, 63, 149-159.

Geisser, S., & Greenhouse, S. W. An extension of Box's results on the use of the F distribution in multivariate analysis. Annals of Mathematical Statistics, 1958, 29, 885-891.

Goldwater, B. C. Psychological significance of pupillary movements. Psychological Bulletin, 1972, 77, 340-355.

Graham, D. T. Health, disease, and the mind-body problem: Linguistic parallelism. Psychosomatic Medicine, 1967, 29, 52-71.

Graham, D. T. Psychophysiology and medicine. Psychophysiology, 1971, 8, 121-131.

Graham, F. K., & Clifton, R. K. Heart-rate change as a component of the orienting response. Psychological Bulletin, 1966, 65, 305-320.

Graham, F. K. Analysis of heart rate response curves: A comment on pooled interaction error terms. Psychophysiology, 1970, 7, 485-489.

Hakerem, G., & Sutton, S. Pupillary response at visual threshold. Nature, 1966, 212, 485-486.

Hart, J. T. Recall, recognition, and the memory-monitoring process. Unpublished doctoral dissertation, Stanford University, 1965. (a)

Hart, J. T. Memory and the feeling-of-knowing experience. Journal of Educational Psychology, 1965, 56, 208-216. (b)

Hart, J. T. Methodological note on feeling-of-knowing experiments. Journal of Educational Psychology, 1966, 57, 347-349.

Hart, J. T. Memory and the memory-monitoring process. Journal of Verbal Learning and Verbal Behavior, 1967, 6, 685-691. (a)

Hart, J. T. Second-try recall, recognition, and the memory-monitoring process. Journal of Educational Psychology, 1967, 58, 193-197. (b)

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart and Winston, 1963.

Heath, H. A., & Oken, D. The quantification of "response" to experimental stimuli. Psychosomatic Medicine, 1965, 27, 457-471.

Hess, E. H. Attitude and pupil size. Scientific American, 1965, 212, 46-54.

Hess, E. H. Pupillometric assessment. In J. M. Shlien (Ed.), Research in psychotherapy, Vol. 3. Washington: American Psychological Association, 1968.

Hess, E. H., & Polt, J. M. Pupil size in relation to mental activity during simple problem-solving. Science, 1964, 143, 1190-1192.

Hess, E. H. Pupillometrics: A method of studying mental, emotional, and sensory processes. In N. S. Greenfield & R. A. Sternbach (Eds.), Handbook of Psychophysiology. New York: Holt, 1972.

Hochhaus, L. A response decisional theory of memory judgment rating. Unpublished manuscript, Iowa State University, 1969.

Hochhaus, L. Editing processes in memory. Unpublished doctoral dissertation, Iowa State University, 1970.

Holtzman, W. H. Statistical models for the study of change in the single case. In C. W. Harris (Ed.), Problems in measuring change. Madison: The University of Wisconsin Press, 1963.

Hopkins, R. H., & Atkinson, R. C. First-letter clues in the retrieval of proper names from long-term memory. Psychological Reports, 1968, 23, 851-866.

Horowitz, L. M., White, M. A., & Atwood, D. W. Word fragments as aids to recall: The organization of a word. Journal of Experimental Psychology, 1968, 76, 219-226.

James, W. The principles of psychology, Vol. 1. New York: Holt, 1899.

Johnson, D. A. Pupillary responses during a short-term memory task: Cognitive processing, arousal, or both? Journal of Experimental Psychology, 1971, 90, 311-318.

Kahneman, D., & Beatty, J. Pupil diameter and load on memory. Science, 1966, 154, 1583-1585.

Kahneman, D., & Beatty, J. Pupillary responses in a pitch-discrimination task. Perception & Psychophysics, 1967, 2, 101-105.

Kahneman, D., Beatty, J., & Pollack, I. Perceptual deficit during a mental task. Science, 1967, 157, 218-219.

Kahneman, D., Onuska, L., & Wolman, R. E. Effects of grouping on the pupillary response in a short-term memory task. Quarterly Journal of Experimental Psychology, 1968, 20, 309-311.

Kahneman, D., Peavler, W. S., & Onuska, L. Effects of verbalization and incentive on the pupil response to mental activity. Canadian Journal of Psychology, 1968, 22, 186-196.

Kahneman, D., & Peavler, W. S. Incentive effects and pupillary changes in association learning. Journal of Experimental Psychology, 1969, 79, 312-318.

Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. Pupillary, heart rate, and skin resistance changes during a mental task. Journal of Experimental Psychology, 1969, 79, 164-167.

Kahneman, D., & Wright, P. Changes of pupil size and rehearsal strategies in a short-term memory task. Quarterly Journal of Experimental Psychology, 1971, 23, 187-196.

Kintsch, W. Learning, memory, and conceptual processes. New York: Wiley, 1970.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, Calif.: Wadsworth, 1968.

Lacey, J. I. The evaluation of autonomic responses: Toward a general solution. Annals of the New York Academy of Sciences, 1956, 67, 123-164.

Lehr, D. J., & Bergum, B. O. Note on pupillary adaptation. Perceptual and Motor Skills, 1966, 23, 917-918.

Lincoln Library of Essential Information, The. Buffalo: Frontier Press, 1963.

Lowenstein, O., & Loewenfeld, I. E. The sleep-waking cycle and pupillary activity. Annals of the New York Academy of Sciences, 1964, 117, 142-156.

Lykken, D. T. Range correction applied to heart rate and to GSR data. Psychophysiology, 1972, 9, 373-379.

Mandler, G. Verbal learning. In N. Kogan & M. Wallach (Eds.), New directions in psychology, Vol. 3, New York: Holt, Rinehart & Winston, 1967.

Myers, J. L. Fundamentals of experimental design. Boston: Allyn & Bacon, 1966.

New York Times Encyclopedic Almanac, The. New York: The New York Times Co., 1972.

Norman, D. A. Memory and attention. New York: Wiley, 1969.

Norman, D. A. Remembrance of things past. Technical Report No. 11, CHIP, University of California, San Diego, 1970.

Norman, D. A. Memory, knowledge, and the answering of questions. Technical Report No. 25, CHIP, University of California, San Diego, 1972.

Norman, D. A. & Rumelhart, D. E. A system for perception and memory. In D. A. Norman (Ed.), Models of human memory. New York: Academic Press, 1970.

Nunnally, J. C., Knott, P. D., Duchnowski, A., & Parker, R. Pupillary response as a general measure of activation. Perception & Psychophysics, 1967, 2, 149-155.

Oken, D., & Heath, H. A. The law of initial values: Some further considerations. Psychosomatic Medicine, 1963, 25, 3-12.

Paivio, A. Latency of verbal associations and imagery to noun stimuli as a function of abstractness and generality. Canadian Journal of Psychology, 1966, 20, 378-387.

Paivio, A., & Simpson, H. M. The effect of word abstractness and pleasantness on pupil size during an imagery task. Psychonomic Science, 1966, 5, 55-56.

Paivio, A., & Simpson, H. M. Magnitude and latency of the pupillary response during an imagery task as a function of stimulus abstractness and imagery ability. Psychonomic Science, 1968, 12, 45-46.

Peavler, W. S. Attention, processing load, and pupil size. Unpublished doctoral dissertation, University of Delaware, 1969.

Polt, J. M. Effect of threat of shock pupillary response in a problem-solving situation. Perceptual and Motor Skills, 1970, 31, 587-593.

Reader's digest 1972 almanac and yearbook. Pleasantville, N. Y.: Reader's Digest Association, 1972.

Schaefer, T., Ferguson, J. B., Klein, J. A., & Rawson, E. B. Pupillary responses during mental activities. Psychonomic Science, 1968, 12, 137-138.

Shiffrin, R. M. Memory search. In D. A. Norman (Ed.), Models of human memory. New York: Academic Press, 1970.

Shiffrin, R. M., & Atkinson, R. C. Storage and retrieval processes in long-term memory, Psychological Review, 1969, 76, 179-193.

Simpson, H. M., & Paivio, A. Changes in pupil size during an imagery task without motor response involvement. Psychonomic Science, 1966, 5, 405-406.

Simpson, H. M., Molloy, F. M., Hale, S. M., & Climan, M. H. Latency and magnitude of the pupillary response during an imagery task. Psychonomic Science, 1968, 13, 293-294.

Simpson, H. M., & Paivio, A. Effects on pupil size of manual and verbal indicators of cognitive task fulfillment. Perception & Psychophysics, 1968, 3, 185-190.

Simpson, H. M. Effects of a task-relevant response on pupil size. Psychophysiology, 1969, 6, 115-121.

Simpson, H. M., & Hale, S. M. Pupillary changes during a decision-making task. Perceptual and Motor Skills, 1969, 29, 495-498.

Simpson, H. M., & Climan, M. H. Pupillary and electromyographic changes during an imagery task. Psychophysiology, 1971, 8, 483-490.

Simpson, H. M., & Molloy, F. M. Effects of audience anxiety on pupil size. Psychophysiology, 1971, 8, 491-496.

Stanners, R. F., & Headley, D. B. Pupil size and instructional set in recognition and recall. Psychophysiology, 1972, 9, 505-511.

Stanners, R. F., Headley, D. B., & Clark, W. R. The pupillary response to sentences: Influences of listening set and deep structure. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 257-263.

Steinschneider, A., & Lipton, E. L. Individual differences in autonomic responsivity: Problems of measurement. Psychosomatic Medicine, 1965, 27, 446-456.

Stern, J. A. Toward a definition of psychophysiology. Psychophysiology, 1964, 1, 90-91.

Sternbach, R. A. Principles of psychophysiology. New York: Academic Press, 1966.

Stoyva, J., & Kamiya, J. Electrophysiological studies of dreaming as the prototype of a new strategy in the study of consciousness. Psychological Review, 1968, 75, 192-205.

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

Tulving, E., & Madigan, S. A. Memory and verbal learning. Annual Review of Psychology, 1970, 21, 437-484.

Underwood, B. J. Attributes of memory. Psychological Review, 1969, 76, 559-573.

Van Egeren, L. F., Headrick, M. W., & Hein, P. L. Individual differences in autonomic responses: Illustration of a possible solution. Psychophysiology, 1972, 9, 626-633.

Wilder, J. Modern psychophysiology and the law of initial value. American Journal of Psychotherapy, 1958, 12, 199-221.

Wilder, J. Stimulus and response: The law of initial value. Bristol, England: John Wright & Sons, 1967.

Wilson, R. S. Analysis of autonomic reaction patterns. Psychophysiology, 1967, 4, 125-142.

Winer, B. J. Statistical principles in experimental design. (2nd Ed.) New York: McGraw-Hill, 1962.

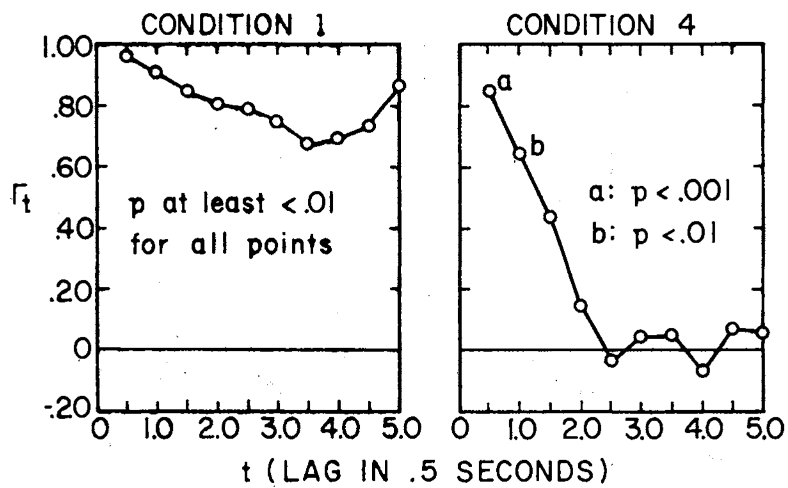
Woodmansee, J. J. Methodological problems in pupillographic experiments. Proceedings of the 74th Annual Convention of the American Psychological Association, 1966, 1, 133-134.

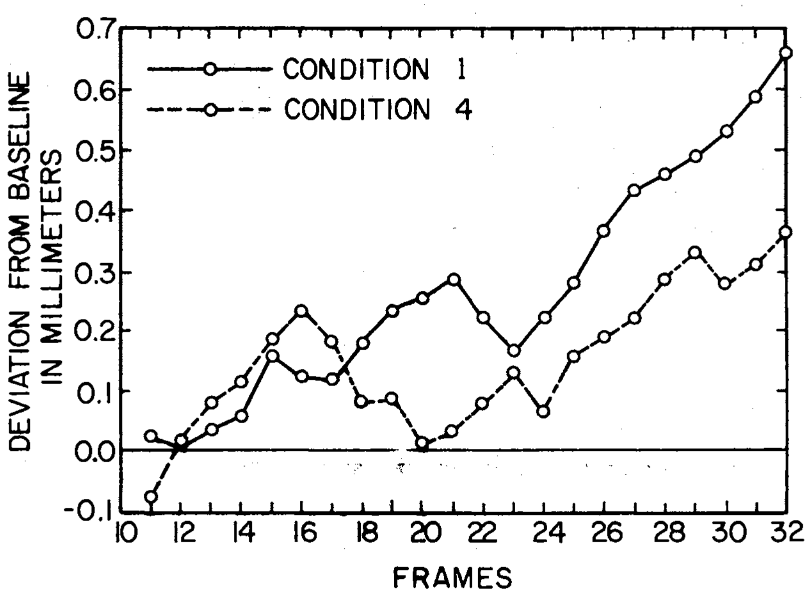
Woodworth, R. S. & Schlosberg, H. Experimental psychology. New York: Holt, Rinehart & Winston, 1954.

Wright, P., & Kahneman, D. Evidence for alternative strategies of sentence retention. Quarterly Journal of Experimental Psychology, 1971, 23, 197-213.

APPENDIX A

EXAMPLE OF SERIAL CORRELATION: CORRELOGRAMS AND TIME-SERIES FOR SELECTED PUPILLOMETRIC DATA





APPENDIX B

QUESTIONNAIRE

The following questions were presented to all Ss in the given order. The first five questions were used as practice trials.

What is the capital of Japan .... Tokyo

Who first sailed around the world .... Ferdinand Magellan

What caverns are in New Mexico .... Carlsbad

Who wrote The House of Seven Gables .... Nathaniel Hawthorne

Who invented the cotton gin .... Eli Whitney

In what country is the Taj Mahal .... India

In what state is Fort Knox .... Kentucky

Who wrote The Grapes of Wrath .... John Steinbeck

What is the capital of Finland .... Helsinki

Who discovered the laws of heredity .... Gregor Mendel

What President is on a 10-cent coin .... Franklin Roosevelt

What is the capital of Portugal .... Lisbon

Who wrote A Tale of Two Cities .... Charles Dickens

What is the largest city in Scotland .... Glasgow

What animal runs the fastest .... Cheetah

Who wrote Poor Richard's Almanac .... Ben Franklin

What is the capital of New Hampshire .... Concord

In what country is Monte Carlo .... Monaco

What instrument records earthquakes .... Seismograph

Who wrote The Three Musketeers .... Alexander Dumas

What is the capital of Austria .... Vienna

Who invented the phonograph .... Thomas Edison

Who writes the comic strip "Peanuts" .... Charles Schulz

In what state is Yale University .... Connecticut

What sea borders Albania .... Adriatic

Who wrote Gulliver's Travels .... Jonathan Swift

What is the capital of Delaware .... Dover

Who is the mayor of San Francisco .... Joseph Alioto

Who wrote Moby Dick .... Herman Melville

What is the state tree of California .... Redwood

Who discovered the North Pole .... Robert Peary

What is another name for the Netherlands .... Holland

Who wrote The Red Badge of Courage .... Stephen Crane

In what city is the Parthenon .... Athens

What lizard can change colors .... Chameleon

Who wrote Paradise Lost .... John Milton

What is the capital of Hungary .... Budapest

What is the longest river in Canada .... Mackenzie

Who wrote Robinson Crusoe .... Daniel Defoe

In what country is the Black Forest .... Germany

What is the capital of Idaho .... Boise

Who invented the steamboat .... Robert Fulton

In what city is the river Thames .... London

Who wrote The Origin of the Species .... Charles Darwin

Who is the "Father of Medicine" .... Hippocrates

APPENDIX C

INSTRUCTIONS TO SUBJECTS

The following instructions were tape recorded and played to all Ss.

This is an experiment concerned with the way in which people remember past knowledge. It is not an intelligence test, and should not be interpreted as such. Also, there is no electric shock or any other kind of unpleasant stimulus. Although your task may seem to be a very simple one, our research indicates that it can provide important information on the way in which people remember things. Therefore, your very close cooperation is absolutely necessary for the success of the experiment. If for any reason during the course of the experiment you feel that you cannot fully cooperate, please let the experimenter know.

As the instructions continue, your attention will be called at various times to information on index cards which will be given to you by the experimenter. These cards are to help you understand the task.

In the experiment, you will be asked a series of questions which involve general knowledge. You will hear the questions over the headphones. The questions are short and specific, and ask for one-word answers. An example of a question is: "What is the capital of Denmark?" Note that there is only one correct answer. This feature characterizes all the questions. Some of the questions you will be able to answer readily. If so, you are to write what you think is the correct answer on a slip of paper, and hand it to the experimenter. If the answer calls for a person's name, give the last name. In the previous example, the correct answer is Copenhagen. If your answer is correct, the green light on the small box by your right hand will light up. If your answer is incorrect, the red light will come on. If you cannot actually give an answer, you will signify whether you think you "know" the answer or not. That is, after some questions, you may not be able to produce the answer, but you may have the feeling nevertheless that you do know it, that it is on the tip of your tongue, so to speak. Thus, you may feel that you can almost "see" the answer in your "mind's eye", but you cannot quite produce it yet; for the moment it is temporarily blocked out of your memory. If this is the case, you will turn the switch by your right hand in the direction labeled "Know". If the question is one that you don't know the answer to, move the switch in the direction marked "Do Not Know". After you have made a rating, move the switch back to the middle position.

If you did not give an answer to the question, or if your answer was incorrect, a copy of the question will be placed above the switch. Then you will be given certain clues about the answer. These clues are the first and last letters of the answer. They will be given in this order, and you will simply hear the two letters. In the previous example, you would hear: "C", "N", which represent the first and last letters of Copenhagen. If after you hear these clues, you are now able to give an answer, write it down, and hand it to the experimenter. If your answer is correct, the green light will come on, and we will proceed to a new question. If you still don't know the answer, you will rate whether you feel you are getting "closer" to it. Sometimes, the clues may start you on the right tract, such that you feel you are on the verge of getting the answer. In this case, move the lever to the position marked "Closer". If you feel that you are no closer to the answer than before, move the lever to "Not Closer".

If you have not recalled the answer by now, it will be presented to you. If the answer involves the name of a person, the first name will be given first, followed by the last name. The clues you were given would refer to the last name. In the present example, you would hear the word "Copenhagen". After you hear the answer, you are to rate whether you recognize it as being the obvious response to the question. For example, many times we try to remember the name of someone we've met before. We feel that if someone were to say a list of names to us, we would be able to reject the incorrect ones, but could recognize the right one. When we hear the correct name, often we respond with something like "Of course", or "That's it!" Thus, if you recognize the answer in this context, move the switch to the "Recognize" direction. Most of the answers to the questions you have probably come across before, but a movement of the switch to this position indicates that you realize that the answer is the obvious response to the question. Otherwise, move the lever to the "Not Recognize" position, which indicates that the answer means nothing to you in relation to the question. All the questions and answers are legitimate ones, and there is no attempt at deception. Any time you are given clues, or an answer, you can be sure that the information is correct. Also, a given answer is used only once. A new question will be presented if you recall the correct answer, or after the answer is presented to you.

There are a few additional procedures for you to follow. When the experimenter is ready to start a trial, the white light in front of you will light up. You should then put your chin on the yellow chin-rest, and fixate your eyes on the cross in the middle of the screen. You will then hear the word "Ready", indicating that the question will be presented in a few seconds. About five seconds after the question is presented, you will hear the word "Respond". At this time you should lean back from the apparatus, and make a response. Either write an answer down, or make a rating by moving the switch.

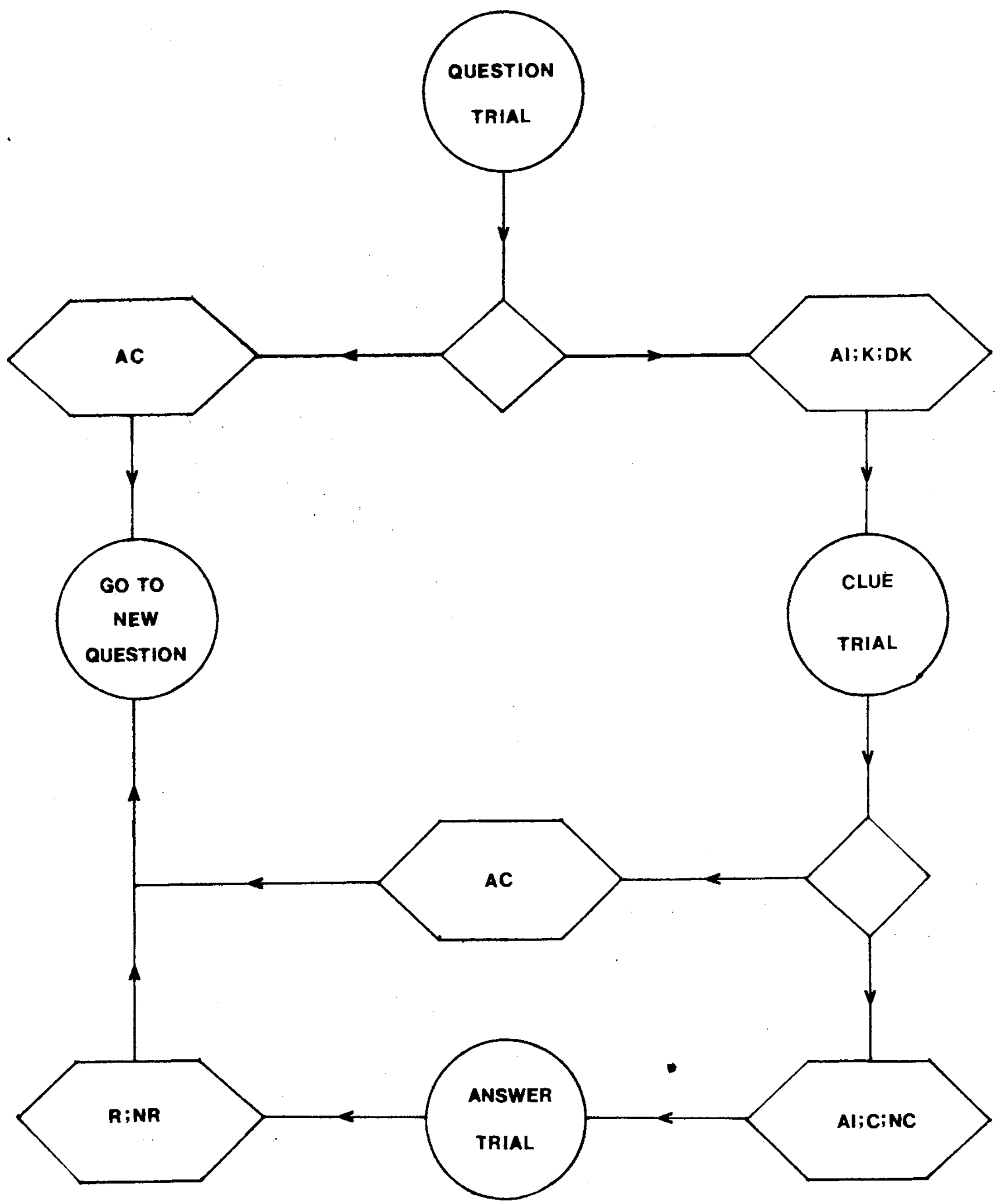
Don't do any writing, though, unless you are going to give an answer. Note that you should do your thinking during the period following the question, so that you will be able to give an answer or make a rating when you hear the word "Respond". Also, until you hear the word "Respond", it is important that you are fixating on the cross, because your eyes will be photographed. During a trial, you will hear a faint background tone. It is simply for apparatus control and you need not pay attention to it. When the experimenter is ready to begin another trial, the white light will come on.

The procedures are the same for a clue, or an answer trial. If after either the question or clue trial, you think you have the correct answer, write it down. You need not be concerned about the correct spelling. Don't hold an answer back if you think you have a good idea of what the correct answer is.

Again it should be emphasized that this task in no way constitutes an intelligence test. Don't be concerned if you miss many questions. The experiment's purpose is to gather information on the way people recall knowledge. Therefore, simply pay attention to what you hear through the headphones. Also, whenever you do not give an answer, it is important that you try to give an accurate rating. Note that you don't have to memorize each question, because if you are not able to supply the answer, a copy of the question will be placed above the switch for your reference during the following phase or phases for that particular question. There will be a few practice trials to get you acquainted with the task. You may now ask any questions you may have.

APPENDIX D

FLOW CHART OF METHODOLOGICAL EVENTS



VITA

Donald Bruce Headley

Candidate for the Degree of

Doctor of Philosophy

Thesis: THE PUPILLARY RESPONSE TO RETRIEVAL OF INFORMATION FROM LONG-TERM MEMORY

Major Field: Psychology

Biographical:

Personal Data: Born in Pittsfield, Massachusetts, August, 1946, the son of Mr. and Mrs. Francis B. Headley.

Education: Graduated from Agawam High School, Agawam, Massachusetts, in June, 1964; received the Bachelor of Science degree from the University of Massachusetts in 1968 with a major in Psychology; received the Master of Science degree from Oklahoma State University, with a major in Psychology, in May, 1971; completed the requirements for the Doctor of Philosophy degree, with a major in Psychology, at Oklahoma State University in July, 1973.

Professional Experience: Served as a Graduate Research Assistant, College of Arts and Science, Oklahoma State University, 1968-1970; was a N.S.F. Summer Research Assistant, 1969-1970; N.S.F. Fellow, 1970-1973; was in charge of Correspondence Introductory Psychology from August, 1970 to August, 1973.