https://ntrs.nasa.gov/search.jsp?R=19670025103 2020-03-12T11:10:17+00:00Z

SPELLING DRILLS

USING A COMPUTER-ASSISTED INSTRUCTIONAL SYSTEM

BY

JACK M. KNUTSON

TECHNICAL REPORT NO. 112

June 30, 1967

NGR-05-020-036

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

N67-34432
(ACCESSION NUMBER)
(THRU)
81
(PAGES)
1
(CODE)
CR-88482
(NASA CR OR TMX OR AD NUMBER)
05
(CATEGORY)
FACILITY FORM 602

TECHNICAL REPORTS

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

(Place of publication shown in parentheses; if published title is different from title of Technical Report, this is also shown in parentheses.)

1 D. Davidson, S. Siegel, and P. Suppes. Some experiments and related theory on the measurement of utility and subjective probability. August 15, 1955. (Experimental test of the basic model, Chapter 2 in Decision-making: An Experimental Approach. Stanford Univ. Press, 1957)

2 P. Suppes. Note on computing all optimal solutions of a dual linear programming problem. November 15, 1955.

3 D. Davidson and P. Suppes. Experimental measurement of utility by use of a linear programming model. April 2, 1956. (Experimental test of a linear programming model, Chapter 3 in Decision-making: An Experimental Approach. Stanford Univ. Press, 1957)

4 E. W. Adams and R. Fagot. A model of riskless choice. August 7, 1956. (Behavioral Science, 1959, 4, 1-10)

5 R. C. Atkinson. A comparison of three models for a Humphreys-type conditioning situation. November 20, 1956.

6 D. Scott and P. Suppes. Foundational aspects of theories of measurement. April 1, 1957. (J. Symbolic Logic, 1958, 23, 113-128)

7 M. Gerlach. Interval measurement of subjective magnitudes with subliminal differences. April 17, 1957.

8 R. C. Atkinson and P. Suppes. An analysis of two-person game situations in terms of statistical learning theory. April 25, 1957. (J. exp. Psychol., 1958, 55, 369-378)

9 R. C. Atkinson and P. Suppes. An analysis of a two-person interaction situation in terms of a Markov process. May 29, 1957. (In R. R. Bush and W. K. Estes (Eds.), Studies in Mathematical Learning Theory. Stanford Univ. Press, 1959. Pp. 65-75)

10 J. Popper and R. C. Atkinson. Discrimination learning in a verbal conditioning situation. July 15, 1957. (J. exp. Psychol., 1958, 56, 21-26)

11 P. Suppes and K. Walsh. A non-linear model for the experimental measurement of utility. August 21, 1956. (Behavioral Science, 1959, 4, 204-211)

12 E. Adams and S. Messick. An axiomatization of Thurstone's successive intervals and paired comparisons scaling models. September 9, 1957. (An axiomatic formulation and generalization of successive intervals scaling, Psychometrika, 1958, 23, 355-368)

13 R. Fagot. An ordered metric model of individual choice behavior. September 12, 1957. (A model for ordered metric scaling by comparison of intervals. Psychometrika, 1959, 24, 157-168)

14 H. Royden, P. Suppes, and K. Walsh. A model for the experimental measurement of the utility of gambling. September 25, 1957. (Behavioral Science, 1959, 4, 11-18)

15 P. Suppes. Two formal models for moral principles. November 1, 1957.

16 W. K. Estes and P. Suppes. Foundations of statistical learning theory, I. The linear model for simple learning. November 20, 1957. (Foundations of linear models. In R. R. Bush and W. K. Estes (Eds.), Studies in Mathematical Learning Theory. Stanford Univ. Press, 1959. Pp. 137-179)

17 D. Davidson and J. Marshak. Experimental tests of a stochastic decision theory. July 25, 1958. (In C. W. Churchman and P. Ratoosh (Eds.), Measurement: Definition and Theories. New York: Wiley, 1959. Pp. 233-269)

18 J. Lamperti and P. Suppes. Chains of infinite order and their application to learning theory. October 15, 1958. (Pacific Journal of Mathematics, 1959, 9, 739-754)

19 P. Suppes. A linear learning model for a continuum of responses. October 18, 1958. (In R. R. Bush and W. K. Estes (Eds.), Studies in Mathematical Learning Theory. Stanford Univ. Press, 1959. Pp. 400-414)

20 P. Suppes. Measurement, empirical meaningfulness and three-valued logic. December 29, 1958. (In C. West Churchman and P. Ratoosh (Eds.), Measurement: Definition and Theories. New York: Wiley, 1959. Pp. 129-143)

21 P. Suppes and R. C. Atkinson. Markov learning models for multiperson situations, I. The theory. February 20, 1959. (Chapter 1 in Markov Learning Models for Multiperson Interaction. Stanford Univ. Press, 1960)

22 J. Lamperti and P. Suppes. Some asymptotic properties of Luce's beta learning model. April 24, 1959. (Psychometrika, 1960, 25, 233-241)

23 P. Suppes. Behavioristic foundations of utility. July 27, 1959. (Econometrica, 1961, 29, 186-202)

24 P. Suppes and F. Krasne. Application of stimulus sampling theory to situations involving social pressure. September 10, 1959. (Psychol. Rev., 1961, 68, 46-59)

25 P. Suppes. Stimulus sampling theory for a continuum of responses. September 11, 1959. (In K. Arrow, S. Karlin, and P. Suppes (Eds.), Mathematical Methods in the Social Sciences. Stanford Univ. Press, 1960. Pp. 348-365)

26 W. K. Estes and P. Suppes. Foundations of statistical learning theory, II. The stimulus sampling model. October 22, 1959.

27 P. Suppes and R. C. Atkinson. Markov learning models for multiperson situations, II. Methods of analysis. December 28, 1959. (Chapter 2 in Markov Learning Models for Multiperson Interactions. Stanford Univ. Press, 1960)

28 R. C. Atkinson. The use of models in experimental psychology. May 24, 1960. (Synthese, 1960, 12, 162-171)

29 R. C. Atkinson. A generalization of stimulus sampling theory. June 14, 1960. (Psychometrika, 1961, 26, 281-290)

30 P. Suppes and J. M. Carlsmith. Experimental analysis of a duopoly situation from the standpoint of mathematical learning theory. June 17, 1960. (International Economic Review, 1962, 3, 1-19)

31 G. Bower. Properties of the one-element model as applied to paired-associate learning. June 29, 1960. (Application of a model to paired-associate learning, Psychometrika, 1961, 26, 255-280)

32 J. H. Blau. The combining of classes condition in learning theory. August 23, 1960. (See Transformation of probabilities, Proceedings of the Amer. Math. Soc., 1961, 12, 511-518)

33 P. Suppes. A comparison of the meaning and uses of models in mathematics and the empirical sciences. August 25, 1960. (Synthese, 1960, 12, 287-301)

34 P. Suppes and J. Zinnes. Stochastic learning theories for a response continuum with non-determinate reinforcement. October 25, 1960. (Psychometrika, 1961, 26, 373-390)

35 P. Suppes and R. Ginsberg. Application of a stimulus sampling model to children's concept formation of binary numbers, with and without an overt correction response. December 14, 1960. (Application of a stimulus sampling model to children's concept formation with and without an overt correction response, Journal exp. Psychol, 1962, 63, 330-336)

(Continued on inside back cover)

SPELLING DRILLS
USING A COMPUTER-ASSISTED INSTRUCTIONAL SYSTEM

by

Jack M. Knutson

TECHNICAL REPORT NO. 112

June 30, 1967

PSYCHOLOGY SERIES

Reproduction in Whole or in Part is Permitted for

any Purpose of the United States Government

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

CHAPTER I

INTRODUCTION AND BACKGROUND OF PROBLEM

During the past few years we have seen an ever-increasing number of books and articles reporting studies in which a digital computer has been used to control part or all of the selection, sequencing and evaluation of instructional materials or lessons, and the students' interactions with these lessons. In general, the term computer-assisted-instruction (CAI) has gained widespread acceptance as a label to refer to an instructional procedure which utilizes a computer in this capacity.

No attempt will be made here to review all of the above-mentioned literature on CAI, for it ranges widely in terms of specificity of curriculum materials used, educational development of students involved, and in the scope of the instructional and learning processes studied. The interested reader may obtain some sense of the diversity and extent of this work through perusal of some of the following publications: Coulson, 1962; Glaser, 1965; Atkinson and Hansen, 1966; Suppes, Hyman, and Jerman, 1966; Suppes, 1964; Suppes, 1967.

While much of the work cited above has consisted of short-term laboratory studies, many universities are presently committed to the study of CAI on a long-term basis and in the context of a more normal classroom situation. Among these one finds the Universities of Illinois, Texas, Pennsylvania State, Florida State, California at Santa Barbara and Irvine, and Stanford.

CHAPTER I

INTRODUCTION AND BACKGROUND OF PROBLEM

During the past few years we have seen an ever-increasing number of books and articles reporting studies in which a digital computer has been used to control part or all of the selection, sequencing and evaluation of instructional materials or lessons, and the students' interactions with these lessons. In general, the term computer-assisted-instruction (CAI) has gained widespread acceptance as a label to refer to an instructional procedure which utilizes a computer in this capacity.

No attempt will be made here to review all of the above-mentioned literature on CAI, for it ranges widely in terms of specificity of curriculum materials used, educational development of students involved, and in the scope of the instructional and learning processes studied. The interested reader may obtain some sense of the diversity and extent of this work through perusal of some of the following publications: Coulson, 1962; Glaser, 1965; Atkinson and Hansen, 1966; Suppes, Hyman, and Jerman, 1966; Suppes, 1964; Suppes, 1967.

While much of the work cited above has consisted of short-term laboratory studies, many universities are presently committed to the study of CAI on a long-term basis and in the context of a more normal classroom situation. Among these one finds the Universities of Illinois, Texas, Pennsylvania State, Florida State, California at Santa Barbara and Irvine, and Stanford.

At the latter school, the Institute for Mathematical Studies in the Social Sciences (IMSSS) has been developing over the last four years a working CAI system for regular classroom usage. This development has used two distinct approaches, which Suppes (1966) refers to as "tutorial systems" and "drill and practice systems."

The tutorial approach to CAI uses the computer in the capacity of "teacher" to present new materials as well as to control subsequent student interactions with them. In this capacity an attempt is being made to teach reading and mathematics to primary-grade children in a school in East Palo Alto. The computer and elaborate terminal equipment being used were developed especially for this purpose by IBM and are located in a separate, new building on the school grounds. A more complete description of this project may be found in several publications (e.g., Atkinson and Hansen, 1966; Wilson and Atkinson, 1967). The important distinction is that a "tutorial" CAI system is designed to approximate or simulate the teacher's normal role and therefore to assume a principal role in the instructional process for part of each school day.

In contrast, drill and practice systems are intended to supplement the instruction which occurs in the classroom. They are designed to improve--through practice--the skills and concepts which are introduced by the classroom teacher. At the same time, this more limited objective makes it possible to use simpler, less expensive equipment.

Beginning in the fall of 1965, CAI drill and practice programs were initiated in two different schools. In both cases a computer at Stanford was hooked up by telephone lines to control teletypes located in the schools. In one school, fourth, fifth and sixth grade students received daily drills in arithmetic (Suppes, Jerman, and Groen, 1965). At the other school, sixth grade children were given daily drills in spelling. Starting in the fall of 1966 this operation was expanded, and currently computer-controlled drills are being given to approximately 800 students in six schools in five different local communities. In addition, an elementary school in a remote area of Kentucky has been linked to the system, and 60 children there are receiving daily drills in arithmetic. This study made use of the equipment and students in the school which has been involved in drill and practice in spelling.

A General Strategy for Research on Drill and Practice in Spelling. It should be emphasized that the research to be reported here is a small part of a total on-going investigation of the potential use and value of CAI drill and practice systems. Hence, it would seem wise to outline briefly the general plan which has been developed to explore some of the problems in spelling. The strategy, as it has evolved, is built on the following considerations.

(1) Relevant experimental studies in the area of verbal learning and spelling do not, for the most part, provide easily interpretable information which may be directly applied to the drill and practice routines in spelling on CAI. For example, given the literally hundreds of studies which have been done on the effects of massed vs. distributed practice on verbal learning, it is exceedingly difficult to decide just which of these findings are applicable when it comes to constructing optimal drill routines for spelling. Rather, the principal value of

the prior research lies in suggesting routines which may then be tested in this new, applied situation.

(2) Relevant variables affecting the rate of learning in drills--once identified--should be explored in a systematic manner through a series of related, short-term studies.

(3) Because of their importance as suggested by earlier pilot studies and experimentation, the first variables to be investigated will be: (a) massed and distributed practice on items; (b) session length as it affects learning; (c) variation in method and portion of stimulus presented; (d) overlearning and its effect on retention of words.

(4) The program of research, while directed toward the exploration of the above-mentioned parameters, shall remain flexible and free to move in new directions if intermediate findings so indicate.

<u>The Use of Drills in Teaching Spelling and Their Possible Relationship to the Spelling Process</u>. Even though there has been a large amount of research on spelling and methods of teaching it, (see Horn, 1960, for a comprehensive review and bibliography), there appears to have been little systematic work done on the role of drills and memory processes in spelling. This deficiency seems particularly unusual when one examines that research which has investigated the efficacy of teaching "rules" or "phonics" techniques for generating the spelling of a word (e.g., Sartorius, 1931; Gates, 1935; Beltramo, 1954; Hahn, 1964). These investigators report varying degrees of success for their techniques of teaching students the phonemic-graphemic regularities in English spelling. They all, however, would seem to agree that the irregularities in the spelling of many English words place severe limitations on this as an exclusive approach to spelling instruction.

A recent study relating to this topic is reported by Hanna, Hanna, Hodges and Rudorf (1966). Hanna and his associates attempted to write an algorithm for a computer program which would generate the correct spelling of a word from the coded phonemes making up the word. Their algorithm to specify the phonemic-to-graphemic correspondences in English spelling is three pages in length, is dependent on the precise dictionary pronunciation of the phonemes, and results in the correct spelling of a word for only fifty percent of the items attempted.

This comment is in no way intended as a criticism of their work; nor is there any implication that knowledge of the spelling regularities specified in their algorithm would not be extremely useful to the individual speller. What seems apparent, however, is that there are a large number of frequently-used words in our language whose spelling must, in part at least, be memorized in a rote fashion.

Certainly, any controvery over the appropriate method to teach spelling would be greatly reduced if we had more precise information concerning the spelling process itself. However, very little is written, and even less seems to be known, about the cognitive processes which are involved in spelling a word. The Hanna study is one of the few places one may find an explicit hypothesis concerning the characteristics of some of these processes. They propose that learning to spell is primarily a cognitive process, and like most learning can be thought of as the building up of strategies or "programs" for organizing and processing information. The good speller then is one who somehow has constructed an efficient and accurate cognitive program for encoding oral speech patterns into their proper graphemic representations. They further recommend that these encoding strategies are best taught inductively, allowing the child to discover for himself that basic structural properties underlie the spellings of many words.

Such information processing theories of learning and behavior are currently quite popular and will perhaps someday produce important knowledge and understanding. It should, however, be made clear that from a "programming" point of view--to continue the analogy--one faces something of a dilemma in constructing a good speller. An experienced computer programmer knows that it is often more efficient, and sometimes necessary, in coming up with the correct solution to a question, to use the machine's memory files rather than some general algorithm to obtain needed information. In this case, where the algorithm produces only fifty percent correct responses, it would seem absolutely essential.

It should be noted here that the Hanna study does not contend that their particular algorithm in any way resembles the actual cognitive strategies which an individual uses in spelling a word. Indeed, they would perhaps argue that the competent speller had developed a "program" which was much more complex and accurate than theirs. The following conceptualization of the spelling process represents an attempt to extend the information processing notion, but at the same time to place greater emphasis on the function of the memory capacity of the individual.

The spelling process could perhaps be thought of as depending on two parallel, interrelated memory processes or strategies for retrieval of information stored in long-term memory. It is conjectured that the individual, when called upon to spell a word, will first of all search his long-term memory store and attempt to find some sort of complete representation of that word. For the mature, competent speller this search seems to be exceedingly rapid and results in the almost instantaneous production of the correct orthography. The processing time required for the good speller to generate most frequently-used words would seem to preclude the possibility that he uses a strategy dependent on internalized rules or algorithms for matching appropriate grapheme to phoneme.

Rather, the second strategy would seem to be called into use mainly when the word is unfamiliar and infrequently used, or when its excessive length for some reason does not permit it to be stored as a single unit. Often when this strategy is apparently being used to spell a word, the individual will, as he proceeds serially through the word, continue to search through his memory for a match between his spelling and his long-term store of graphemic representations of words and sounds.

In addition, it would appear likely that during the learning phase of this spelling process, both of these retrieval strategies should be developed concurrently if we wish to maximize efficiency and accuracy. Since the first strategy is dependent on the size and accessibility of the individual's well-learned word store, an instructional system which focused exclusively on teaching rules and phonemic-graphemic regularities would seem to be inadequate. We must also attempt to strengthen and increase the individual's store of frequently-used words and their associated spelling. Hopefully, a drill and practice routine using a CAI system provides an efficient method of accomplishing this objective.

While it is a matter of conjecture how the associations between words and their correct spellings are established, it seems plausible to think of them as being learned in a manner similar to that in which a paired-associate item is acquired. If such an assumption is reasonable, and the learning of the correct spelling of a word does in some way resemble the learning of a paired-associate item, several potential research questions are immediately apparent. One of the more interesting was suggested in a recent article by Greeno (1964) which reported a paired-associate experiment that perhaps had implications for spelling.

Greeno compared two practice conditions in which the repetitions of some items were distributed in the normal sequence and others were repeated on successive trials. He found evidence indicating that little or no learning occurred on the second trial when an item was repreated immediately or very soon after a previous presentation. In addition, he found that the distributed condition produced more learning than the massed condition, and argued that this result provided evidence for a discrimination theory of paired-associate learning as opposed to a reinforcement or contiguity theory.

One could contend that learning to spell a word involves a similar discrimination process. For example, suppose a child is called on to spell the world "BOAT." Assume also that the child has had sufficient experience with the language and spelling so that the response he generates is not simply a random series of letters. If he is not sure of the word, he is likely to consider a number of possible spellings--for example, BOT, BOTE, and BOAT. Now in order to spell the word correctly, he must learn to discriminate among several--and in this case, reasonable--alternatives. Eventually we would hope that he would eliminate the wrong alternatives and associate just the correct spelling with the spoken word.

If this description does resemble the process that actually occurs in learning to spell a word, then we would expect that greater learning would occur when practice on unlearned words in a list is distributed rather than massed.

Directly relevant to this problem are two unpublished studies (Keller, 1966; Fishman, 1967) which were run using the same CAI system used in this experiment. In his study, Keller presented words under two conditions. Words in the first condition, if missed, were corrected and the next item presented. Words in the second condition were not corrected immediately; rather, the student was informed that he was wrong and told to try again. If after the second try he still misspelled the word, he was given the correct spelling and told to copy it. In spite of the greater time and number of practices which the students had on the second condition words, the probability of a correct response on these was not significantly better than for those words which were not repeated. It would seem likely that one of the principal reasons for this unusual finding was that the words were too easy (overall probability of an error on the retention test was less than 10%). Thus the potential effects of the extra practice on condition two words may have been obscured because most of the students could easily learn the words they did not know without the benefit of extra practice trials.

Fishman's study attempted to evaluate the results of presenting words in massed or distributed trials. Two groups of three words each were presented once every other day over a period of six days. The learning trials on four other groups of words were massed so that all of the trials for that group occurred on the same day. She found that the probability of a correct response for the words in the massed conditions was higher than those distributed during the trial sessions, but that on two-week retention tests the words learned under distributed practice were remembered at a higher rate.

These studies indicate that we do not yet understand clearly the effects of varying our method of dealing with incorrect responses, or of the optimal routines for spacing practices on an item. The present experiment was designed to further explore both of these problems.

Explicitly, both Greeno's and Keller's studies provide evidence suggesting that the immediate repetition of error items will produce minimal learning on the repeated presentations of an item. At the same time, both Greeno's and Fishman's results indicate that distributed spacing of practice trials on error items would be superior. In this study it was decided to investigate the issue by presenting words under three conditions. When a word was misspelled, it was either (1) not repeated during that session, (2) repeated immediately, or (3) repeated after four other items had been presented. If Greeno's findings are applicable in this situation, then we might expect the first two conditions to produce approximately equal learning, i.e., immediate repetition of an item would not greatly increase learning for that item. On the other hand, the spaced practice on items in the third condition should be consistently superior in affecting rate of learning.

Individual Confidence in Spelling. Implicit in the earlier discussion on the possible nature of the spelling process is the notion that as the maturing speller is augmenting his store of quickly-retrievable words and acquiring strategies for generating the regularly-spelled words, he must at the same time develop a third necessary skill. No matter what combination of processes he may use to generate the spelling of a word, he must make decisions as he progresses in his response as to the correctness of a particular spelling. Presumably upon completion of his response, if not before, he must decide whether or not his spelling can be matched with whatever internal representation he may have of the word. In most cases the experienced speller will be able to generate a response which in some way approximates the correct spelling, but he may or may not fail when he tries to make an accurate assessment of the adequacy of his answer.

The reasons for such a failure are difficult to specify exactly. For example, a person might decide that his response is correct because he has an inaccurate representation of the word in his long-term memory. Or he may have generated the spelling by using inappropriate strategies dependent on phonemic-graphemic regularities. Most important, the decision process itself at the juncture when one decides on the correctness of the response may be inadequate for several reasons.

For example, an individual may base his decision on something other than a systematic attempt to match his response with a correct representation in his memory. He might, for instance, base his decision on a generalized self-perception that he is a "good" or "bad" speller. He could also be influenced by the immediate--and perhaps irrelevant--context in which he is asked to respond. Thus, if he had just been told he had misspelled three words in a row, he would perhaps display a tendency to decide that his next spelling also was wrong.

In spite of these and related problems, it was decided that this decision process was such an important component of spelling behavior that it would be interesting to obtain a measure of the student's confidence in the correctness of his particular spelling of a word. Such a measure, it was hoped, would provide partial answers to the following questions: (1) How accurate are elementary school children at assessing the correctness of their spelling? and (2) Are they capable of discriminating subjective impressions of their accuracy into several categories?

While a number of studies involving college-age students (e.g., Bernbach, 1966; Atkinson and Shiffrin, 1967) have reported obtaining a measure of a person's confidence in his response, we are not aware of any research using this dependent variable with elementary school children.

Summary of Purposes of the Study. To summarize briefly, the principal purposes of the study were twofold. First of all, we wished to examine the effects of varying the number and spacing of repetitions on error items to see how these factors would influence the rate of learning of new spelling words in a drill context. Therefore, words were presented under one of three conditions: (1) no repetition of error items during a session; (2) immediate repetition of error items;

(3) repetition of error items after four other words had been presented. Our second major purpose was to test whether or not a student's confidence in the correctness of his particular spelling of a word was reliably related to his performance in spelling that word.

Our overall general purpose was to continue the long-term investigation of the feasibility of using a CAI system of drill and practice as both a pedagogical and research tool. For example, specific problems relating to programming, computer hardware, and data collection and storage are not fully resolved. Similarly, questions concerning coordination of effort between the university research team and the public school staff are constantly arising. Studies such as this will hopefully contribute valuable information leading to the eventual solution of many of these issues.

CHAPTER II

DESIGN AND EXPERIMENTAL PROCEDURE

Subjects. The subjects used in this study were the children in two sixth grade classes in an elementary school in East Palo Alto. Originally sixty students began the experiment, but due to transfers to other schools, absenteeism, and related problems, only forty-two children completed their lists and the subsequent retention tests.

The school which the subjects attend is located in a "culturally disadvantaged" area, and the vast majority of the students are non-Caucasian. As measured on the Lorge-Thorndike Intelligence Test, the mean I.Q. for those subjects who completed the experiment was 85; the range of scores was from 60 to 101, with only two subjects scoring above 100. It should be noted that this is a group-administered test which is largely verbal in nature, and which almost certainly underestimates the real ability of the students.

In reading achievement the students also appear to be below average. Using the Stanford Achievement Tests for intermediate grades, the mean reading achievement score for the group was just over the minimum 4th grade level, which would place them approximately two years below their actual sixth grade placement and well into the bottom quartile of the national averages.

A question may legitimately be raised as to why such an atypical sample of elementary school children was chosen for the study. First of all, it was felt that in part the students' below-average achievement in school reflected the cumulative effects of a lack of basic language-arts skills which could perhaps be improved through the kind of drills which CAI can provide. Secondly, it was hypothesized that the uniqueness and newness of a CAI system may be particularly motivating for a group of students who are typically less interested in school achievement than their middle-class counterparts.

Equipment. The computer used in this experiment is a modified Digital Equipment Corporation PDP-1, with a variety of input and output devices. Essentially, this machine is a high speed digital computer which has been programmed to service a maximum of 28 user programs simultaneously on a real-time sharing basis. The time-sharing capability is made possible through the use of a high speed memory drum which reads user programs in and out of the computer's core memory at a rapid rate.

The audio system which was used is made up of a small Westinghouse P-50 computer which controls twelve tape drives. This smaller computer is linked directly to the PDP-1 and is controlled by programs running on the latter. Each of the tape drives can randomly access any desired segment of a continuous loop of magnetic tape in well under two seconds. These tape loops are 24 inches in circumference, six inches wide, and contain 128 separate tracks. On each track eight one-second segments may be recorded and accessed. Both the PDP-1 and the audio system were connected directly by telephone lines to the terminal equipment at the school.

The terminal equipment was set up in a converted storeroom located a short distance from the two classrooms. In the storeroom were located four student stations, each containing a standard Model-33 Teletype and a set of earphones. Each station also contained an extra audio outlet so that all audio messages could be monitored without disturbing the students as they proceeded through a lesson. To minimize distractions and noise from other teletypes, these stations were separated from each other by four-foot high room-dividers.

All four terminals were controlled by a single program on the PDP-1; each student was serviced sequentially in a round-robin cycle. Due to the extremely rapid speed of the computer, the student received the impression that he was getting "full-time" service, although actually the computer devoted only a small fraction of its running time to any one individual.

Preliminary Training and Orientation. In order to provide a thorough introduction to the experimental procedures and to accustom the students to working on a teletype, an extensive orientation and training program was run during the fall of 1966. Also, since the controlling computer program and part of the terminal equipment had been added to the system in the late summer, it proved to be an opportune time to correct some of the technical and hardware problems which inevitably arise in the development of new CAI systems.

After meeting with the whole class and explaining in general terms the procedure to be followed on the spelling drills, the experimenter and his assistant demonstrated and explained the use of the teletype and audio system to each student individually over a two-week period.

During this introductory phase all students were run on the same list of "easy" fourth grade words. The principal emphasis at this time was in developing familiarity with the teletype keyboard and the overall administrative procedures involved in leaving the classroom, checking into the terminal room, adjusting the earphones, etc. No attempt was made to teach the students to use standard typing techniques; rather they were instructed and aided in finding the letters on the keyboard and encouraged to be as accurate as possible in their typing. This latter consideration was especially important since there was no way for them to erase or correct a typing error, and the computer program which evaluated their responses did not distinguish between typing and other spelling errors.

After all the students had had an opportunity to go through several practice sessions and were beginning to develop a reasonable facility at finding and typing the letters, the idea of using confidence ratings was introduced. Again the students were instructed individually and given practice until they seemed to understand the use and meaning of the various ratings. From the time when the students were first introduced to the equipment up to this point where they had learned to use confidence ratings, four weeks had elapsed.

The original intention had been to begin the formal experiment at the completion of the practice and training period. However, a major hardware problem developed and the computer system was completely inoperative for an extended period. Rather than beginning the experiment and having it interrupted at the mid-point by Christmas vacation, it was decided to use the time remaining until vacation to give the students further practice and work on word lists somewhat easier than those which they would encounter in the experiment. The other procedures used in these practice sessions were identical to those in the experiment.

Word Selection. Careful study of the students' performance during the early practice sessions, consultation with their teachers, and their ability as exhibited on achievement tests all indicated that our original intention of using uniform lists of words for all students was grossly inadequate and impractical because of the heterogeneity of the classes.

In deciding on the word list to be used in the experiment, four criteria were used. (1) The words should have high frequency of occurrence and should be useful in the students' writing. (2) The words should be appropriately difficult so that the student would miss a minimum of 50% of the words on his first attempt at spelling them. (3) On the other hand, the words should be easy enough so that by his fifth time through a list the student would spell at least 10 out of 12 of these words correctly. (4) The list length should be such that a student could complete an entire drill in a ten-minute session.

To meet all of these criteria, it is obvious that each student would have to use completely individualized lists, specifically tailored to his ability and educational achievement. While such individualization of curriculum materials is perhaps an ultimate aim of CAI, as a more realistic compromise it was decided to use three major levels of word difficulty, to construct lists twelve words in length, and to attempt to place each student on that level which would maximize his rate of learning during a ten-minute drill session.

Construction of Word Lists. The problem of constructing word lists that are of similar difficulty is complex. Many factors--the length of the word, the number of permissible variant spellings for the phonemes in the word, and the frequency of occurrence of the word in English--probably are important in determining spelling difficulty. Individuals will also differ greatly in their language experience and facility, so that a word which is easy for one student to spell might be impossibly hard for another.

Bearing all of these complications in mind, it was decided to use the New Iowa Spelling Scale (Greene, 1954) as the source of an objective measure of the difficulty of a word. This scale is the product of the testing of some 238,000 pupils throughout the country in the early 1950's to determine what percentage of students at each grade level could spell a word correctly. Using this measure, the words for the study were selected in the following manner.

First of all, three general pools of words were formed by listing alphabetically for each of three grade levels--4th, 5th and 6th grades--all of the words in the New Iowa Spelling Scale which were spelled correctly by 40% to 50% of the students in each of these grades. Then for each grade level, six lists of twelve words were formed by (1) randomly selecting six words at a time of equal difficulty from the general grade-level pool, and (2) randomly assigning each word to one of the six lists. Using this procedure, the eighteen lists in Appendix A were constructed. It should be noted that the six lists on each grade level are equivalent in range of difficulty and in the number of words represented by any one percentage measure.

Assignment of List Levels and Experimental Conditions. Upon completion of the construction of the lists as described above, the 60 students were assigned to go through six lists on one of the three grade levels. A student's assignment to a particular level was based on his performance in the practice sessions and his ability as measured on the achievement tests. The criteria for moving to a new list within a level were that the student either (1) spelled 10 out of 12 words correctly on the first try on that item on a particular day, or (2) had gone through the same list for five days without meeting the first criterion.

On all grade levels and word lists, the following daily presentation procedures were used. For a particular student and twelve-item list, four words were assigned to one of three conditions:

1. No Repetition (R0) Words in this condition were not repeated during the session, regardless of the correctness of the subject's response.
2. Immediate Repetition (R1) Words in this condition which were misspelled were repeated immediately.
3. Spaced Repetition (R4) Words in this condition which were misspelled were repeated after the presentation of four intervening items.

Words were never presented more than twice in a session, so that if a subject missed all of the words in R1 and R4, he would have a maximum of 20 presentations for that session--eight of which would be repetitions of error items.

Words, conditions and subjects were balanced so that any particular word was always given under R0 to one-third of the subjects, R1 for the second third, and under R4 for the remainder of the subjects. Table 1 summarizes the general experimental conditions for all lists on all grade levels. On their appropriate grade level, subjects and words were randomly assigned within this pattern. Since each student who completed the experiment went through six lists, the total time for the experiment for a particular student was a maximum of 30 school days. The minimum time, of course, depended on how fast the student learned the list, but it was seldom less than 20 school days.

Daily Operation During Experiment. A full-time monitor was on duty whenever the children were using the teletypes. Her presence was primarily a precautionary measure so that an adult was available in case of an equipment failure or other emergency. The actual check-in, presentation and evaluation of the drill, and the sign-out were all handled by the CAI system and occurred as follows.

The student entered the room, sat down at a free terminal, and put on his earphones. On the page-printer on the teletype he saw the message, "Please type your number." After he typed his identification number and depressed the space bar--the latter operation was used as a termination signal for all student responses--the computer then looked up this number in the student history section of the program, determined what list the student was working on that day, randomly permuted the order of presentation of the individual words on the

TABLE I

General Pattern for Assigning Conditions

to Words and Subjects at Each Grade Level

Word No.	Group 1	Group 2	Group 3
1	R0	R1	R4
2	R1	R4	R0
3	R4	R0	R1
4	R0	R1	R4
5	R1	R4	R0
6	R4	R0	R1
7	R0	R1	R4
8	R1	R4	R0
9	R4	R0	R1
10	R0	R1	R4
11	R1	R4	R0
12	R4	R0	R1

list, and positioned the tape on the audio system. Next it sent a command to the teletype to print the student's name, his list number, and the date and time the session was beginning.

At this point the child heard over the earphones the message, "If you hear the audio, please type an 'a' and a space," and at the same time saw this same message printed on the page. If there were something wrong with the audio system, he would call the monitor. Ordinarily he proceeded by following the spoken directions, i.e., he typed an "a" and a space, and the lesson began.

To signal the student that a word was about to be presented, the machine typed a dash (-). The audio system then presented a word, used the word in a sentence, and then repeated the word again. As soon as the audio track was through playing, the machine typed the number of that particular item (1,2,3,etc.). This was the student's signal to begin his response. When he finished typing his answer, he depressed the space bar. The machine then waited for him to type one of four numerals (1,2,3, or 4) as an indication of his degree of confidence in his answer.

Attached to each machine was a small chart reminding the student of the meaning of each confidence rating. The child was told to type the number which identified the phrase most closely corresponding to his feeling. The phrases were:

(1) Positive word is right

(2) Fairly sure word is right

(3) Fairly sure word is wrong

(4) Positive word is wrong

Immediately after receiving this number the computer evaluated the student's response. If the student had responded correctly, the teletype printed out after the student's spelling of the word the message, "--C--", letting the student know he was right. If he was wrong, it typed the message, "--X--", followed by several spaces and a correct spelling of the word. If for some reason the student had not completed his response after forty seconds, the machine typed out the message, "--TU--", meaning time is up. As on a wrong answer this message was followed by several spaces and the correct spelling of the word.

Following all incorrect or timed-out responses, the student was given six seconds to study the correct answer before the next item was presented. On correct responses the study time was three seconds.

After proceeding through all of the items on the list in a similar fashion, the student received a printed message, "End," followed by his name, list number for the next session, the date and ending time, and the number of words he spelled correctly on the day's lesson. These daily drills were collected by the monitor, and at no time during the experiment was the student given a copy of the words to study on his own.

A flow chart summarizing the presentation procedure may be found in Figure 1.

Retention Tests. Retention tests on all words which an individual had gone through were given one week after the entire group had finished their experimental runs. Since individuals varied in the number of runs they had needed to reach criterion, this meant that the time from the last run on a list until the retention test also varied.

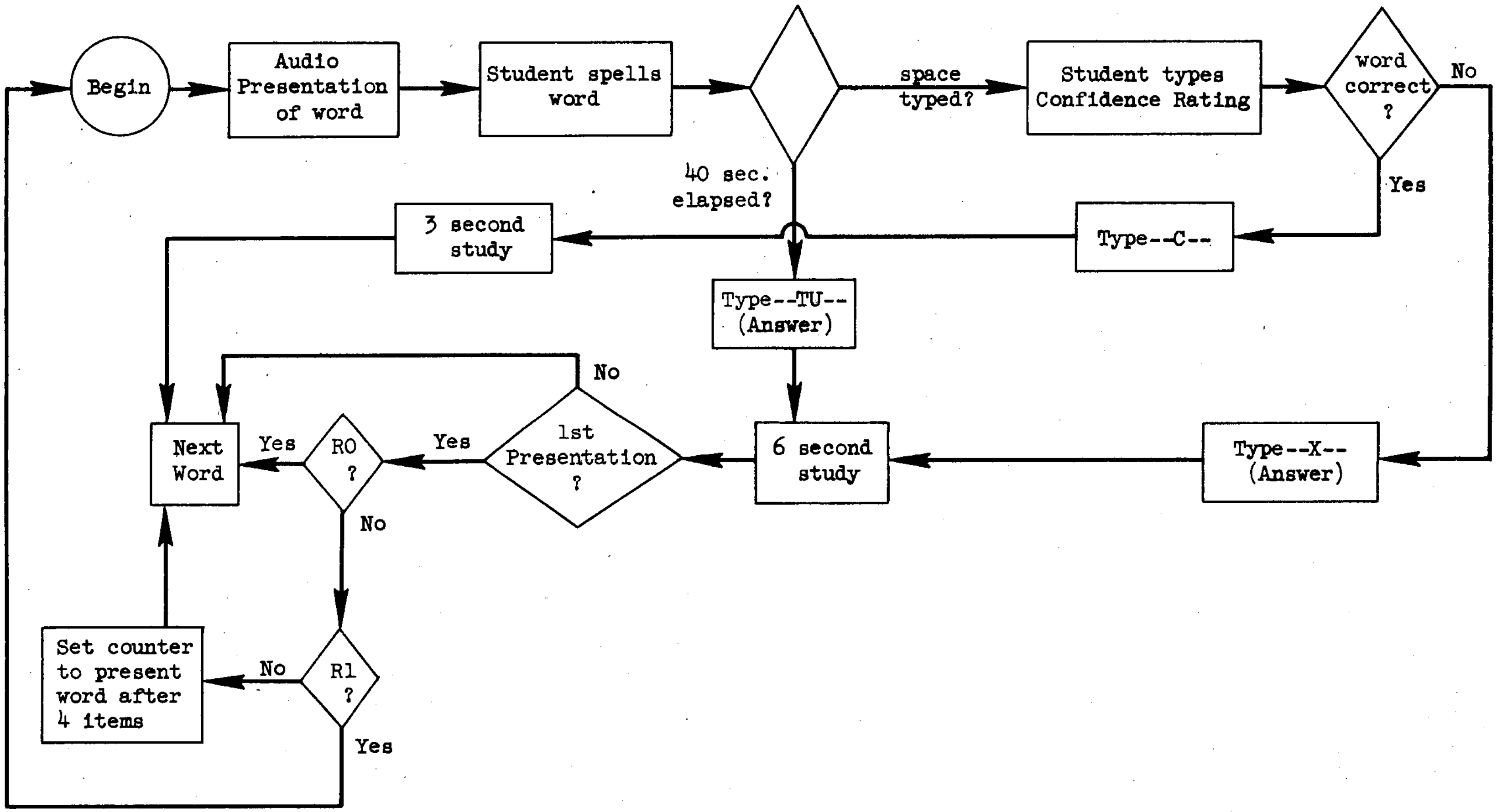

Figure 1. Flow chart for daily presentation routine for each word on 12 item list

In the case of the first list the individual had studied in the experiment, the retention interval was approximately six weeks. For the last list studied, the minimum interval was one week; the average retention interval for the entire group and all lists was just over three weeks.

Retention tests were given in the same manner as the experimental runs, with the exception that error items were not repeated. The students were given immediate feedback about the correctness of their responses.

Since time limitations prohibited testing a subject on all six lists on the same day, the retention tests were divided into two parts. Lists for Weeks 1-3 were given on the first day, and lists for Weeks 4-6 were presented on the succeeding day. (Rather unexpectedly, the fact that these test sessions were somewhat longer than a normal practice session may have affected student performance adversely. This problem will be discussed later in the Results Section.)

Data Collection. For every student response the following information was transmitted to the controlling computer program and stored on disk memory:

1. Subject identification number
2. List number
3. Run number (i.e., was this the first, second, or nth time the student had gone through the list?)
4. Relative item number of the word for that session (i.e., was word presented 1st, 2nd, 3rd, etc.?)
5. Identification number for the word

6. Experimental condition number under which the word was presented

7. Number indicating whether this was the 1st or 2nd presentation of the word for that session

8. Was the response correct, an error, or a time-out?

9. Confidence rating

10. Time in hundredths of a second from the end of the audio message to the first letter of the student response

11. Time in hundredths of a second to the last letter of the student's response

12. Time from the completion of the response to the confidence rating

With the exception of the latency information all of this data was also retrievable from the daily lesson print-outs at the school. One of the jobs of the monitor at the terminal site was to keep a daily tally of the student performance; this information was then compared with the data stored by the computer and served as a valuable means for checking accuracy.

CHAPTER III

ANALYSIS OF EXPERIMENTAL RESULTS

Problem of Incomplete Data. Analysis of the outcome of the experiment was made somewhat complicated by the following considerations: (1) Eighteen of the original sixty students who began the study did not go through all six lists as the original design had specified. (2) Because of the criteria used to advance a student to a new list, the total number of runs for any one subject varied considerably from person to person.

The first problem of incomplete data may be handled in several ways. The common solution to this dilemma seems to be, in most educational research, to use the data from only those subjects completing the entire experiment. From a purely statistical point of view this appears to be a satisfactory procedure, so long as one is very careful about subsequent generalizations to the larger population.

In this experiment it was decided to examine the data in two phases. First, the data from all the subjects--whether or not they had completed all six lists--was summarized and plotted on graphs. This data from the entire original group did not differ substantively from that which was used in the more thorough, subsequent analyses for the smaller group of subjects who completed the entire experiment.

Re-definition of Groups. Of the original 60 subjects, 21 out of 23 students using the 6th grade words completed their lists; 15 out of 18 students on the 5th grade words finished the experiment; in the group using the 4th grade words, only 6 out of 19 students finished. This high attrition rate in the latter group was due to: (1) three students transferred to other schools; (2) four students had to be dropped for disciplinary reasons (e.g., refusal to follow directions); and (3) high rate of absenteeism.

Consequently, for purposes of analysis the remaining subjects were regrouped into two groups, composed as follows:

(1) High Group -- 21 subjects, all of whom ran on 6th grade words;

(2) Low Group -- 21 subjects, 15 ran on 5th grade words; 6 used 4th grade words.

First Analysis--Total Errors to Criterion

The first major analysis was based on the total number of errors a subject made on a particular list from Run 2 up to and including the run on which he reached the 10 out of 12 criterion--or had made five runs without reaching this criterion. Table 2 illustrates the basis for computing the score for each condition on each list. Errors on repeat items for a run are not included in the scoring. As can be seen, a student's score for a particular condition could range from 0 errors to a maximum of 16 if he missed all the items for that condition on every run.

A four-way analysis of variance for a mixed model with fixed and random effects was computed, using as the major dimensions: 2 Groups (High and Low); 3 Conditions (R0, R1, R4); 6 Weeks (1st through 6th); and 21 Individuals (per group).

The dimension labeled Weeks is somewhat misleading, for all individuals did not use six weeks to go through their lists.

TABLE 2

Possible Number of Errors for Each

Condition and List on Runs 2 to 5

Condition	R0	R1	R4
Run 2	4	4	4
Run 3	4	4	4
Run 4	4	4	4
Run 5	4	4	4
Total	16	16	16

The inclusion of this faction in the analysis was prompted by inspection of the total error data (see Figure 3). The observable general decrease in errors as subjects went through their lists suggested that a possible "learning-to-learn" effect was occurring. To test whether or not this could be the case the Weeks dimension was included, and represents the order in which a subject proceeded through his six lists. Since list order was randomized, this decrease in errors does not represent the level of difficulty for a particular list of words.

The complete results of this analysis, including the interactions, are presented in Table 3. The error term for computing the F-ratios for the three main effects (i.e., Groups, Conditions, and Weeks) is not simply the residual mean square. In the case of the Group effect, the error term was found by adding the sum of squares of Individuals,

TABLE 3

Analysis of Variance for Errors to Criterion

Sources of Variation	Degrees of Freedom	Sums of Squares	Mean Squares	F-Ratio	d.f.	Probability
Individuals	20	1972.919	98.646			
Groups	1	953.437	953.437	8.45	1/40	$p < .01$
Conditions	2	237.428	118.714	20.042	2/680	$p < .001$
Weeks	5	262.202	52.440	8.853	5/680	$p < .001$
Individ. X Groups	20	2540.256	127.012			
Individ. X Cond.	40	191.960	4.799			
Individ. X Weeks	100	1002.269	10.022			
Groups X Cond.	2	28.795	14.397	2.43	2/680	$p < .10$
Groups X Weeks	5	71.506	14.301	2.414	5/680	$p < .10$
Cond. X Weeks	10	24.762	2.476			
Ind. X Groups X Cond.	40	213.258	5.331			
Ind. X Groups X Weeks	100	1085.626	10.856			
Ind. X Cond. X Weeks	200	783.511	3.917			
Groups X Weeks X Cond.	10	34.474	3.447			
Residual	200	751.125	3.755			
Total	755	10153.534				

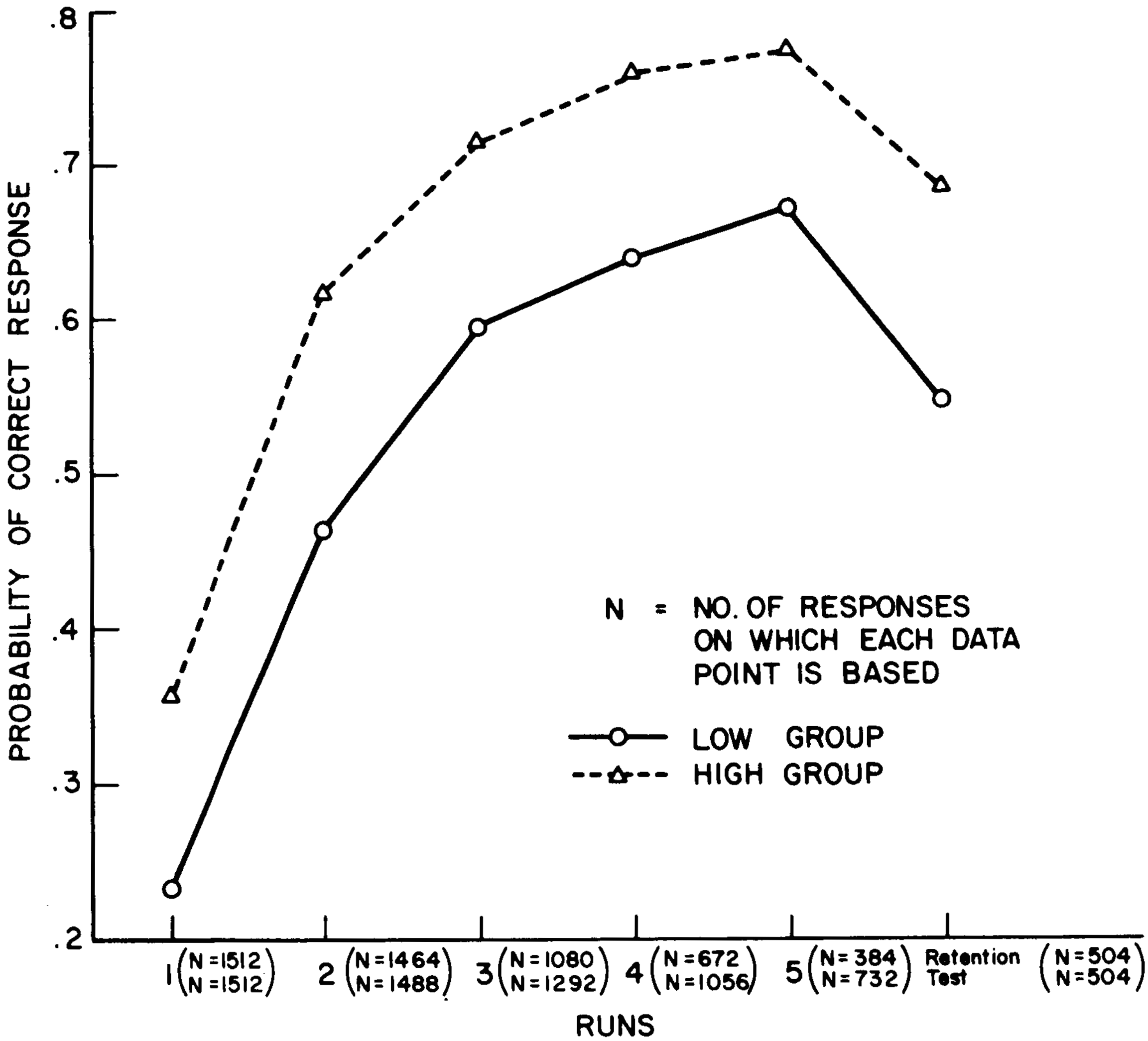

Figure 2. Probability of a correct response over runs for each group

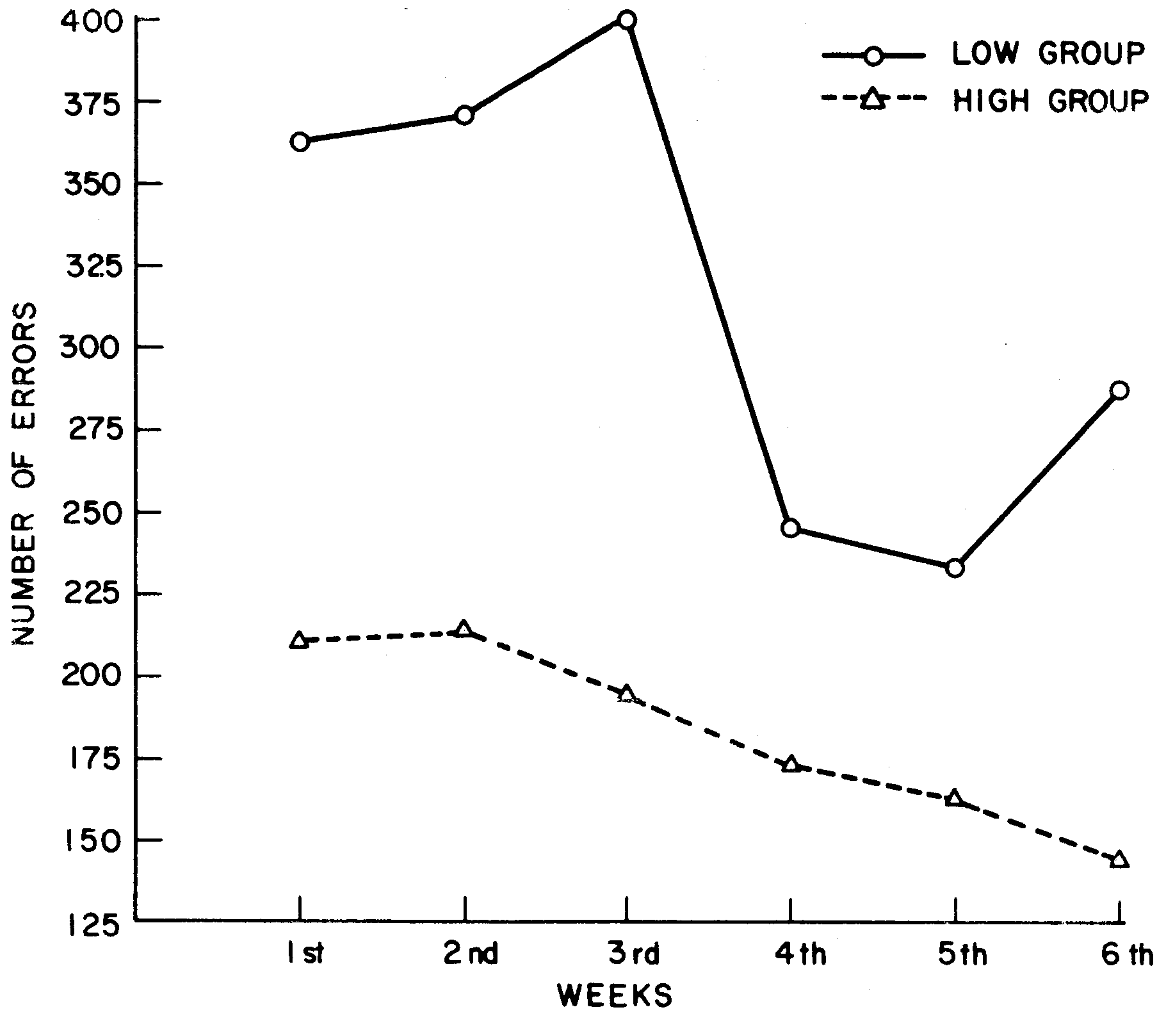

Figure 3. Total number of errors to criterion or five runs on list (Total excludes Run 1 on all lists and includes only first presentation of an item for any one run.)

and the Group by Individuals interaction, and dividing by the total degrees of freedom (i.e., 40). For the other two main effects, the error term is the sum of the following interactions: (1) Groups by Conditions; (2) Groups by Weeks; (3) Individuals by Groups by Conditions; (4) Individuals by Groups by Weeks; (5) Individuals by Conditions by Weeks; and (6) Residual; this total is then divided by the sum of the degrees of freedom (i.e., 680). None of these interactions are significant or particularly interesting since they include variation due to expected individual differences. In any case, this estimate of the error term is certainly conservative in the sense that it increases the size of the error variance, and thus is less likely to produce a significant F-ratio for the main effects.

Group Differences. The significant difference ($F = 8.45$, $p < .01$) found between the Low and High groups was not surprising, since the groups were selected on the basis of their ability and were run on different lists of words. The initial hope had been that the use of harder words for the High group would serve to make the task equally difficult for this group as for the Low group. That the experimenter was not successful in equalizing relative difficulty is readily apparent in Figure 2, which shows the overall learning rate of the two groups as they proceeded over runs on a list.

Week Differences. A highly significant difference ($F = 8.853$, $p < .001$) was found for the Weeks effect. Interestingly the Groups by Weeks interaction approached significance ($F = 2.414$, $p < .10$). This differential effect of Weeks on the two groups--as well as the overall decrease in total errors--may be clearly seen in Figure 3. It should be recalled that this experiment was begun after each subject had had several weeks of orientation and training on the task. Evidently, the complexity of the response demanded--involving as it does motoric skills and memory processes--is such that factors other than knowledge of specific words is being learned.

Condition Differences. The variation in number and spacing of repetitions on error trials produced a significant difference ($F = 20.042$, $p < .001$) between conditions. This difference is reflected in Figures 4 and 5, which show the probability of a correct response on the first try for all items over runs. Inspection of these graphs does not make it immediately clear if the significance found is due to the large difference between R0 (no repetition of error items) and the other two conditions, or if the apparent superiority of R4 over R1 is also statistically significant.

Consequently, t-tests for correlated scores (McNemar, 1962) were run to compare R1 (immediate repetition of an error item) and R4 (repetition of an error item after four intervening items). The results are summarized in Table 4. As may be seen, these conditions do differ significantly for the combined groups and for the Low group, but not for the High. This difference in the effect of the conditions on the Low and High groups is also indicated in the interaction term (Group by Condition) for the analysis of variance, which approaches significance ($F = 2.43$, $p < .10$).

Analysis of Retention Test Scores

An analysis of variance for the number correct for retention test scores was carried out employing the same model and dimensions as those

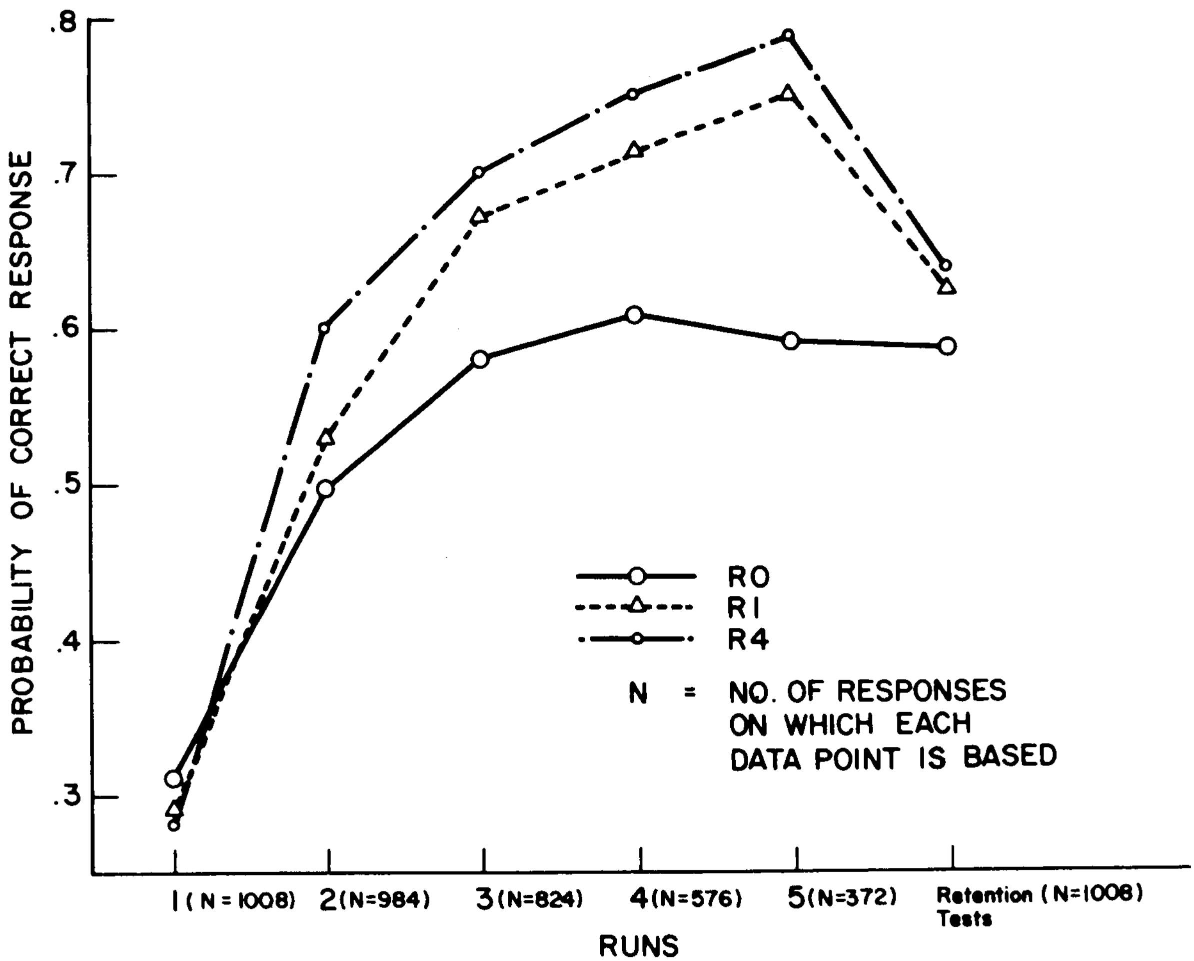

Figure 4. Probability of a correct response by condition for the combined groups over runs and retention test

TABLE 4

T - Tests for Total Errors to Criterion
(R1 vs. R4)

Combined Groups	$t_{251} = 3.19$	$p < .001$
High Group	$t_{125} = .568$	not significant
Low Group	$t_{125} = 3.42$	$p < .001$

used for the total error analysis. However, since a subject went through each list only once on the test, the possible range of scores for each cell was 0 to 4, the latter being the maximum he could have correct for a particular condition on one list.

The error terms used to test for significance were computed in the manner described earlier. The results of this analysis are presented in Table 5.

Group Differences. Again, as might be expected, the test scores for the two groups were significantly different ($F = 9.91$, $p < .005$). As may be seen on Figure 6, the difference in the probability of a correct response for the two groups is approximately .10. When this figure is compared with the initial probability of a correct response on Run 1 (see Figure 2), it may be seen that the relative difference between the two groups is approximately the same in both situations.

Week Differences. The Weeks dimension proved to be significantly different, ($F = 3.34$, $p < .01$), but the type of trend previously observed in Figure 3 is not evident here. Since the retention

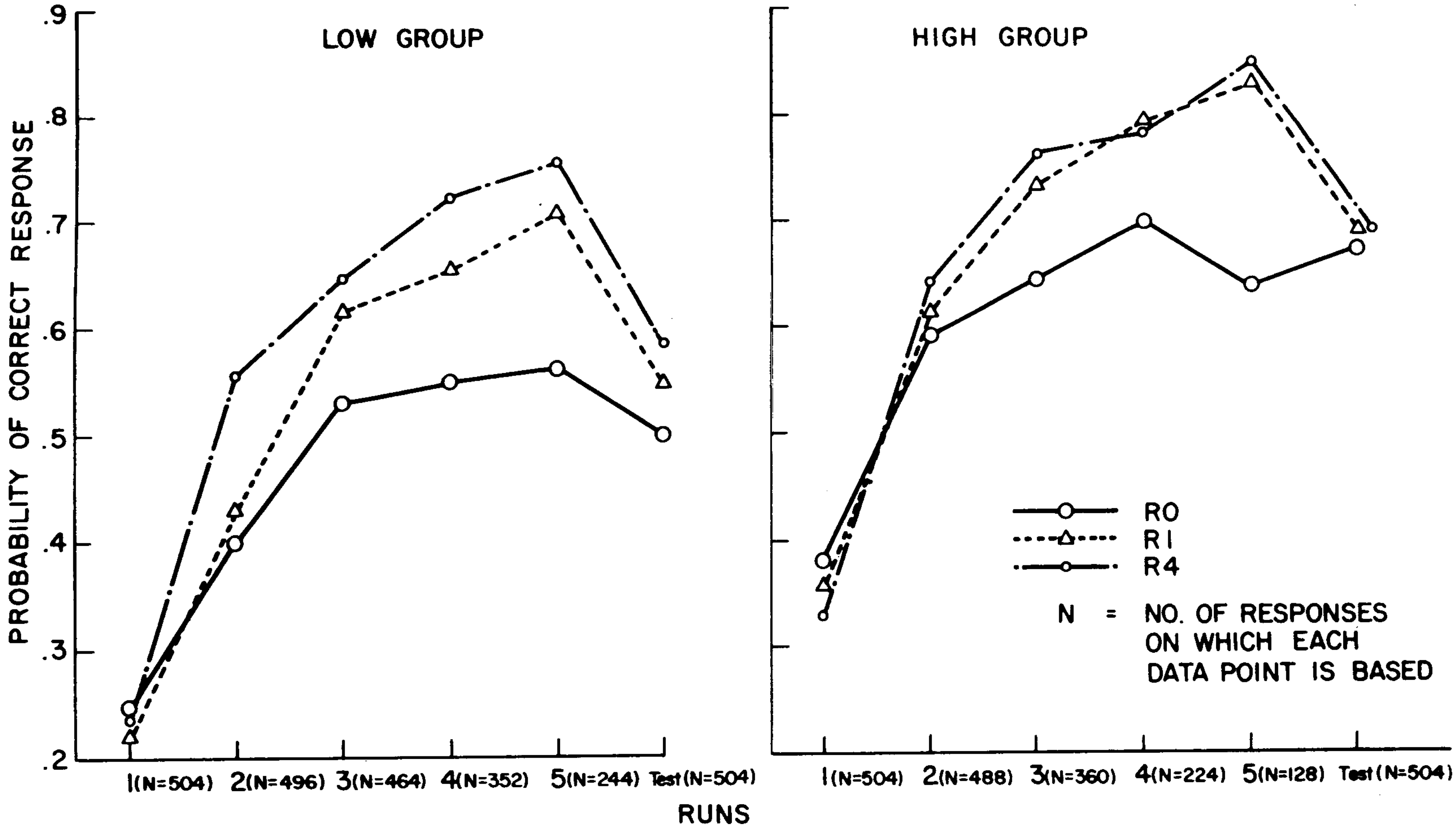

Figure 5. Probability of correct response by condition for each group

TABLE 5

Analysis of Variance for Retention Test Scores

Sources of Variation	Degrees of Freedom	Sums of Squares	Mean Squares	F-Ratio	d.f.	Probability
Individuals	20	82.026	4.101			
Groups	1	61.713	61.713	9.91	1/40	$p < .005$
Conditions	2	5.923	2.961	3.187	2/680	$p < .05$
Weeks	5	15.518	3.103	3.34	5/680	$p < .01$
Indiv. X Groups	20	166.952	8.347			
Indiv. X Cond.	40	38.521	0.963			
Indiv. X Weeks	100	97.925	0.979			
Groups X Cond.	2	2.166	1.083			
Groups X Weeks	5	3.047	0.609			
Cond. X Weeks	10	8.886	0.888			
Ind. X Groups X Cond.	40	51.166	1.279			
Ind. X Groups X Weeks	100	119.951	1.199			
Ind. X Cond. X Weeks	200	177.668	0.888			
Groups X Cond. X Weeks	10	6.356	0.635			
Residual	200	146.642	0.733			
Total	755	948.469				

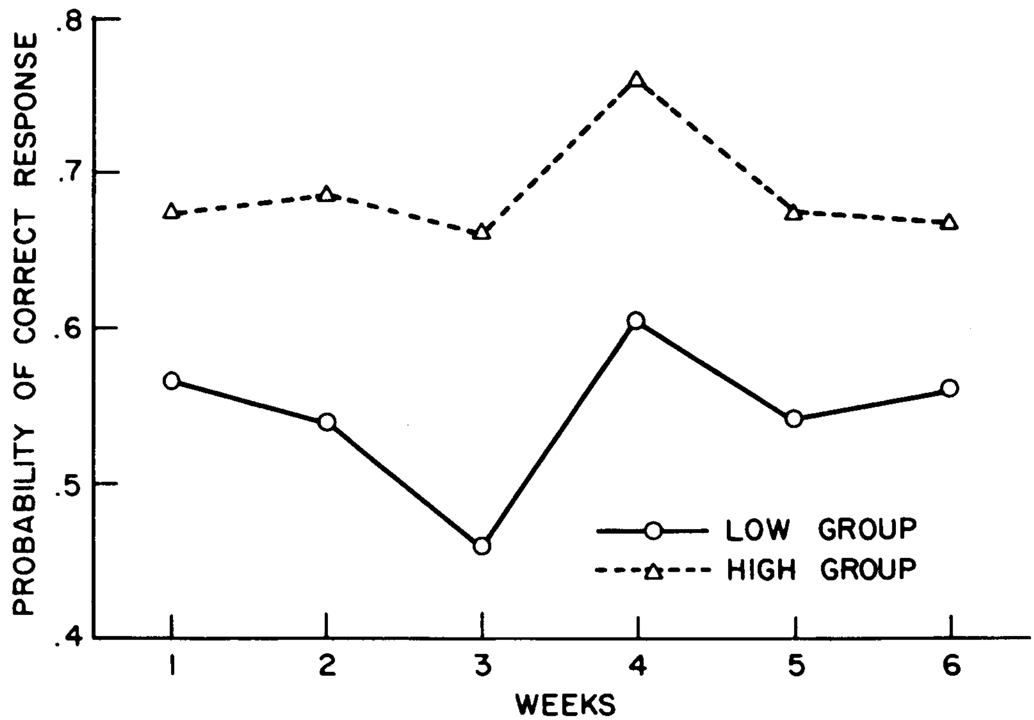

Figure 6. Probability of correct response on retention tests for low and high groups

interval varied considerably for words in Week 1 compared with those in Week 6, for example, one might have reasonably expected a general increase in correct responses over the Weeks dimension. Figure 6, however, seems to indicate that the varying retention interval did not produce any easily interpretable, systematic effects.

The dramatic difference between the 3rd and 4th Week test scores for both groups seemed rather unusual until it was realized that this difference may be primarily attributable to using the scores from retention tests given over a two-day period and to using longer-than-normal sessions for each subject's testing. Thus, the general slight downward trend for Weeks 1 to 3, followed by the sharp rise on Week 4 and another downward trend for the remaining weeks may well be due to some sort of fatigue factor operating within each of the two test sessions. This unfortunate confounding of effects could have been avoided if the words had been randomized over all six lists instead of merely within each list. This was not done because the experimenter assumed the increase in session length would not affect the results so noticeably.

Condition Differences. The condition differences were significant ($F = 3.187$, $p < .05$), but the magnitude of this difference was considerably less than in the test for total errors.

Table 6 presents the results of running a t-test for correlated scores for the various condition combinations. It can be seen that the main significance effects come principally from the Low group, and involve mainly the difference in retention between no-repetition items (R0) and those which were repeated (R1 and R4). However, it should be noted that the trend observed in the first analysis of error scores is

TABLE 6

T - Tests for Condition Differences on Retention Tests

Conditions	Low Group		High Group	
RO vs. R1	$t_{125} = 1.86$	$p < .10$	$t_{125} = .81$	not signif.
RO vs. R4	$t_{125} = 2.68$	$p < .01$	$t_{125} = 1.35$	not signif.
R1 vs. R4	$t_{125} = .96$	not signif.	$t_{125} = 1.23$	not signif

also found here (i.e., the probability of being correct is greatest for R4 followed by R1 and then RO)

Analysis of Difference Scores

The scores for this analysis were computed by substracting a subject's score (number correct) on Run 1 for a list from his score on the retention test for that same list. Since the maximum number correct for each condition on a list is 4, the obtained difference scores could range from -4 (if the subject had all of the words correct on Run 1 and missed all of these items on the retention test) to +4 (if the subject missed all the items on Run 1 and had them all correct on the retention test). Unlike the scores for total errors to criterion, this difference score is not directly affected by the number of runs a subject might have had on a list.

The results of this analysis are found in Table 7. The only source of variation which was significant was that for Conditions ($F = 3.51$, $p < .05$).

TABLE 7

Analysis of Variance for Difference Scores

Sources of Variation	Degrees of Freedom	Sums of Squares	Mean Squares	F-Ratio	d.f.	Probability
Individuals	20	55.740	2.787			
Groups	1	0.894	0.894	0.258	1/40	not signif.
Weeks	5	8.756	1.751	1.096	5/680	not signif.
Conditions	2	12.835	6.417	4.018	2/680	$p < .05$
Ind. X Groups	20	82.661	4.133			
Ind. X Weeks	100	207.798	2.077			
Ind. X Cond.	40	51.052	1.276			
Groups X Weeks	5	15.820	3.164	1.981	5/680	$p < .10$
Groups X Cond.	2	0.645	0.322			
Weeks X Cond.	10	9.608	0.960			
Ind. X Groups X Weeks	100	217.290	2.172			
Ind. X Groups X Cond.	40	66.465	1.661			
Ind. X Week X Cond.	200	251.835	1.259			
Groups X Weeks X Cond.	10	12.211	1.221			
Residual	200	224.009	1.120			
Total	755	1217.627				

TABLE 8

T - Tests for Condition Differences for Difference Scores

Conditions	Low Group		High Group	
R0 vs. R1	$t_{125} = 4.04$	$p < .001$	$t_{125} = 1.18$	not signif.
R0 vs. R4	$t_{125} = 2.31$	$p < .05$	$t_{125} = 1.91$	not signif.
R1 vs. R4	$t_{125} = .63$	not signif.	$t_{125} = .51$	not signif.

The three condition differences were compared using t-tests for correlated scores, and the results are summarized in Table 8.

As was the case on the retention test scores, the effects of the conditions seem to be more pronounced for the Low group. The Group by Condition interaction is not, however, significant ($F = .202$, $p < .90$).

The lack of a significant difference between groups ($F = .258$, $.75 > p > .50$) may at first glance seem unusual until one recalls that what is being tested here is the significance of differences in gain scores. If one looks, for example, at Figure 2, it is easy to see that the overall gain in proportion correct for the two groups is very similar.

The effect along the Weeks variable does not reach significance ($F = 1.096$, $.5 > p > .25$), although the Groups by Weeks interaction is higher ($F = 1.981$, $p < .10$). The same criticism of possible attenuation of test scores because of test-session length may, of course, be leveled here.

Latency Data

In order to insure that the latency to the response was measured from the same beginning point, the students were explicitly instructed not to begin their answer until the number of the item was printed by the teletype. The first latency then was the number of seconds elapsing from the time the item number was printed until the subject struck the first key. The second recorded latency was measured from the same beginning point to the termination response (i.e., depressing the space bar). The third latency was the time from the termination signal to the confidence rating.

This method of obtaining a uniform starting point for the latency measures may mask real differences which exist in the needed processing time for a particular word or condition. It should be recalled that each item is pronounced three times, once in a sentence context and twice alone. The total time for the audio presentation and the typing of the item number was just under five seconds. Once the student is familiar with the word, presumably after the first run, this time during the audio presentation would be available to the student to use as he needs or desires. For example, if the item is one he is uncertain of, he could use the last four seconds of the audio presentation period to begin his recall or encoding of the proper response. On the other hand, if he feels he knows the word, he could use this time to daydream or even to become frustrated with the slowness of the whole procedure. One cannot be sure that such a difference in necessary processing time will be accurately reflected in the latency as measured.

Why then was the completion of the audio message used as the starting point for latency measures? Perhaps, the student should have been allowed to respond as soon as he was able; the first key he struck should have terminated the audio, and the time from the beginning of the word presentation to the first key used as the latency measure. This procedure was not followed for two reasons: (1) The lack of consistent high fidelity and reliability in the transmission of audio messages made it seem necessary to repeat the word to maximize accurate perception of the item; and (2) the controlling computer program would have had to have been modified extensively in order to allow the student to terminate the audio before its completion. Since there was little evidence indicating that such a modification was really necessary or of great importance in this essentially explorative phase of the total study, it was not made.

All of the latency data reported in this section was averaged separately for each individual; thus, each individual is weighted equally in the overall means, and the slower-learning student is not represented disproportionately simply because he took more runs to learn a list.

Latencies for Correct and Error Responses. Figure 7 shows the overall difference in mean latencies between correct and incorrect responses for the combined groups. This difference seems to be quite congruent with most of the experimental literature which generally reports a higher latency for incorrect responses.

On Figure 8 the correct and error latencies are depicted as they diminish over runs on a list. Again, such a difference was expected and is in keeping with general findings.

Latencies for Conditions. The latencies for each condition as they changed over runs on a list is displayed in Figures 9 and 10. Here the results are somewhat more ambiguous, but in general the rather small differences between latencies for the conditions probably reflects the difference between the number of correct and incorrect responses in each condition. For example, in Figure 10 for the Low group, where the R1 and R4 latencies are quite consistently lower than R0, the latencies for the former conditions may be interpreted as a result of the fact that items in R1 and R4 are more likely to be correct.

Latencies to End of Response. Figure 11, which presents the latencies to the last letter of the response, is included as an example of the information obtained from this measure. They too seem to follow the trends apparent in the latency to the first letter, and are quite in line with one's intuitive expectations--i.e., as the student learns the word better, he can complete his response faster.

Confidence Rating Measures

Because of the great individual variations in the use of confidence rating categories, it is rather difficult to report overall summary statistics for this measure which do not distort the results to some degree.

As with the earlier latency data, all means and frequencies were computed separately for each individual and the overall means calculated from these averages.

The overall distribution of the use of the various confidence ratings for the two groups can be seen in Figure 12. The overwhelming tendency of most subjects to be either positive they are right or positive they are wrong appears even more extreme if one inspects the individual distributions for each rating. Here one observes that approximately two-thirds of the subjects used categories 2 and 3 less than 5 percent of the time.

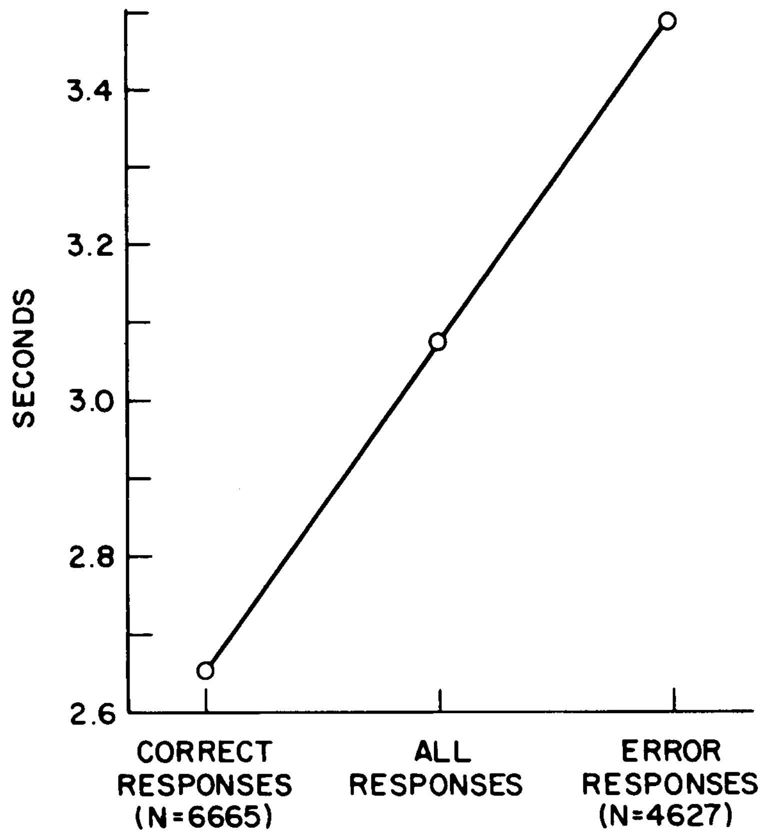

Figure 7. Mean latencies to first letter for correct and error responses for combined groups

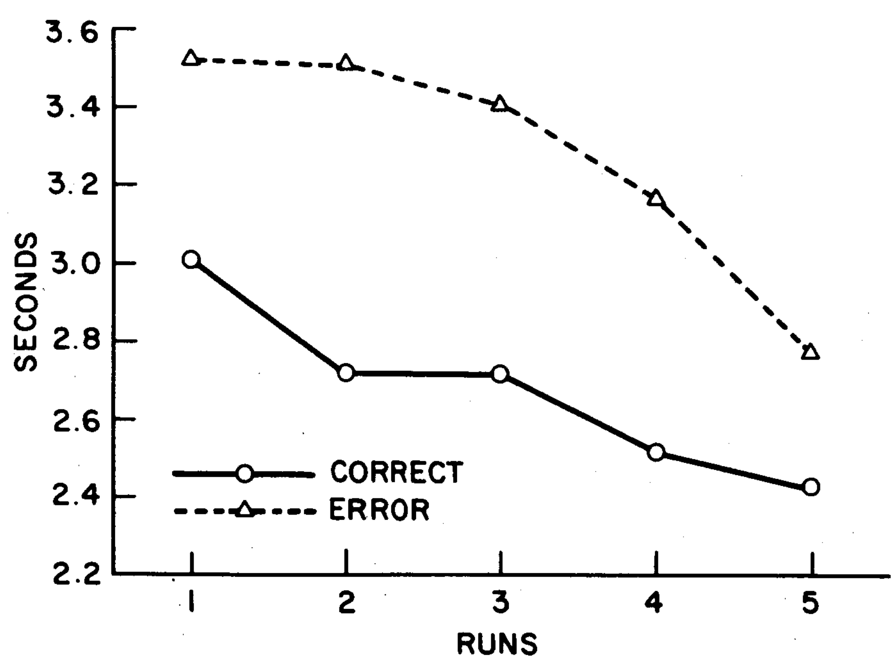

Figure 8. Mean latencies to first letter by runs for correct and error responses for combined groups

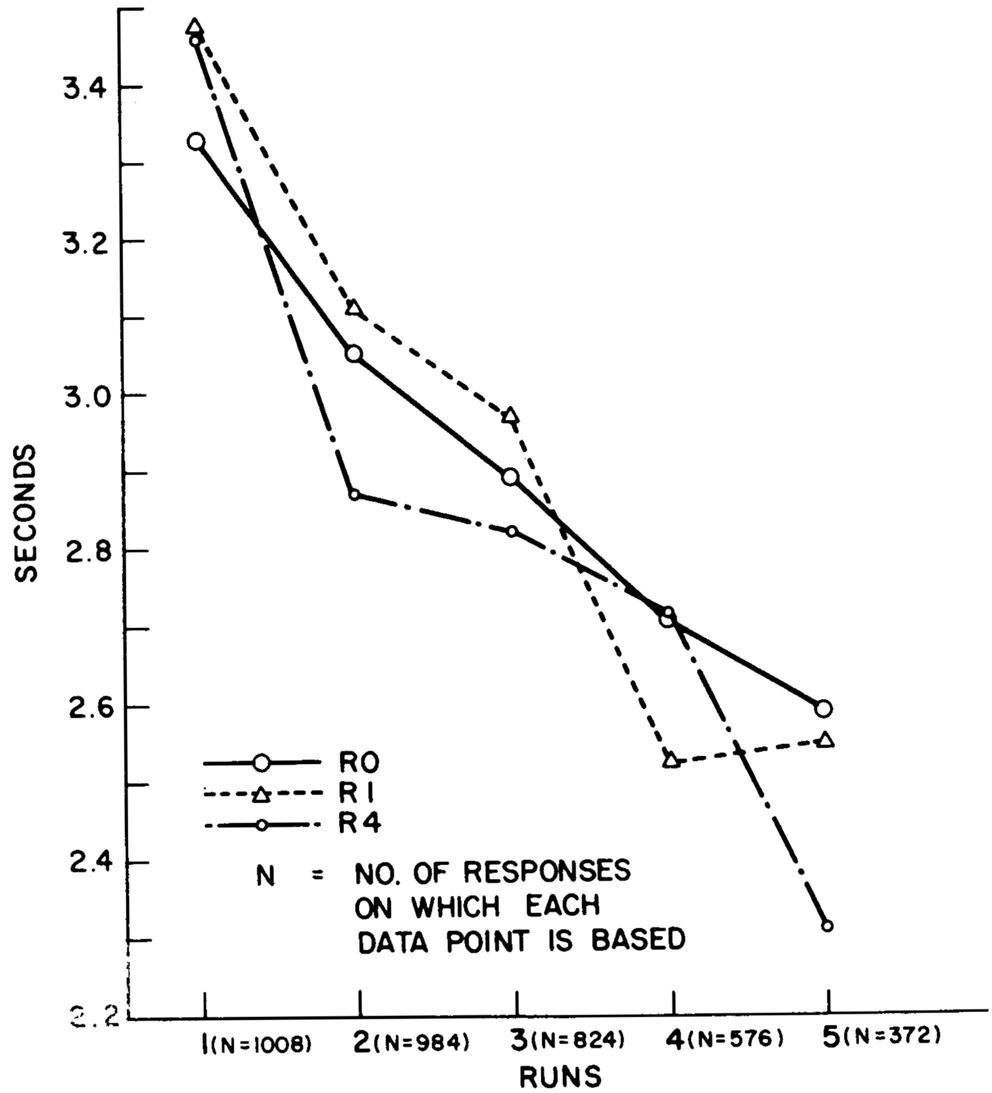

Figure 9. Mean latency to first letter by condition for combined groups

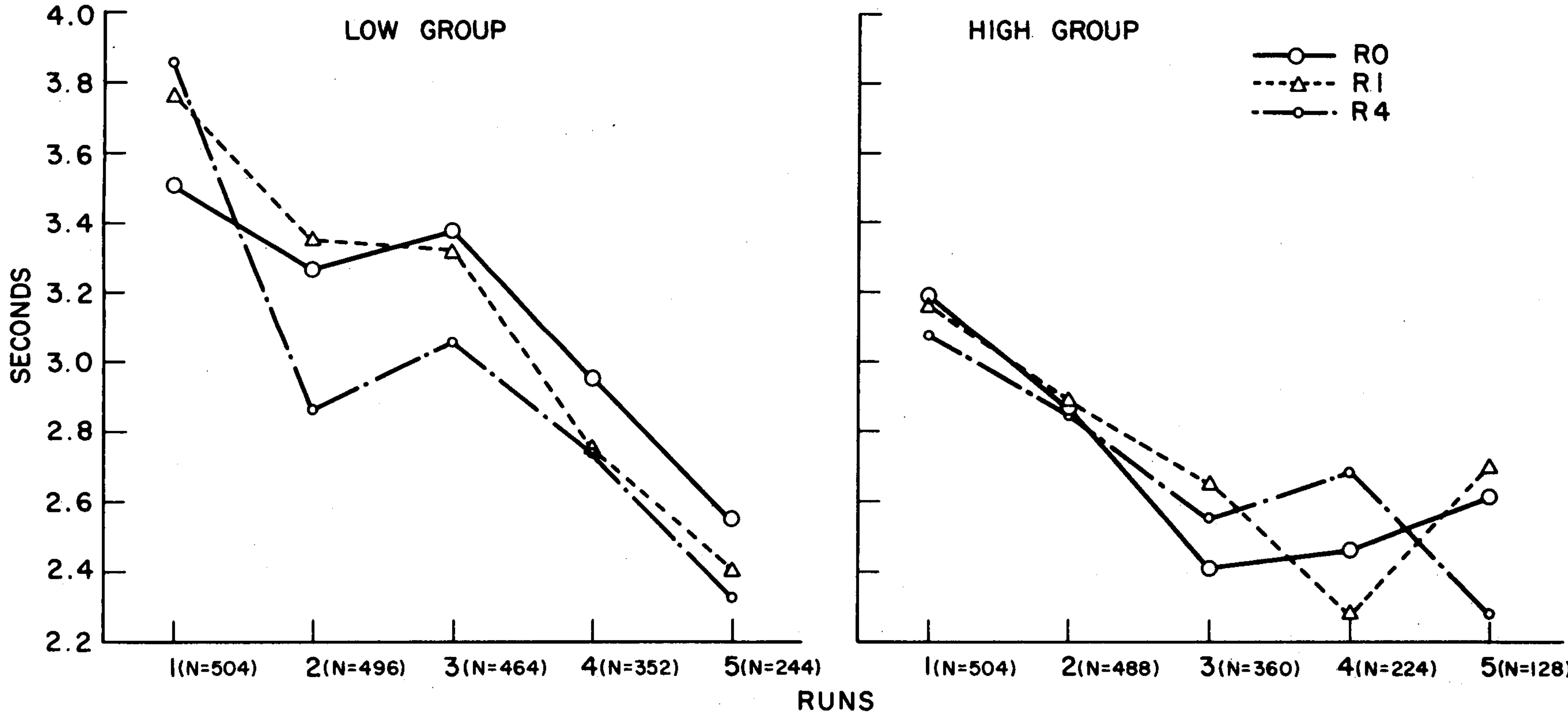

Figure 10. Mean latency to first letter by condition over runs

The low usage of categories 2 and 3 by most subjects does not, however, invalidate the findings reported in Figure 13, where we see the probability of a correct response given the various confidence ratings. In this case also one may observe great individual differences in the accuracy of the ratings, with a relatively few individuals exerting a large influence on the overall averages in categories 2 and 3.

The relationship of the confidence rating to the latency to the first letter of the response is seen in Figure 14. With the possible exception of the latency for confidence rating 2 in the High group, the latency measures seem to reflect the expected rising uncertainty indicated by the confidence rating.

Figure 15 represents the latencies for the confidence ratings (i.e., the time from the termination signal to the time the key for the confidence rating measure was struck).

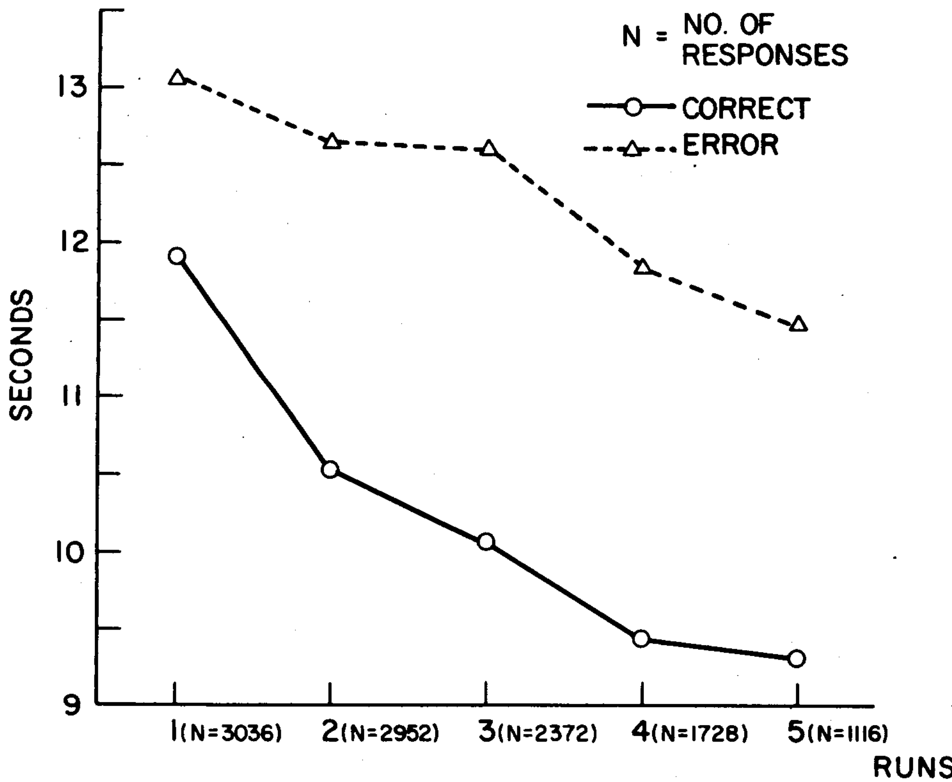

Figure 11a. Mean latency to last letter of response for correct and error over runs for combined groups

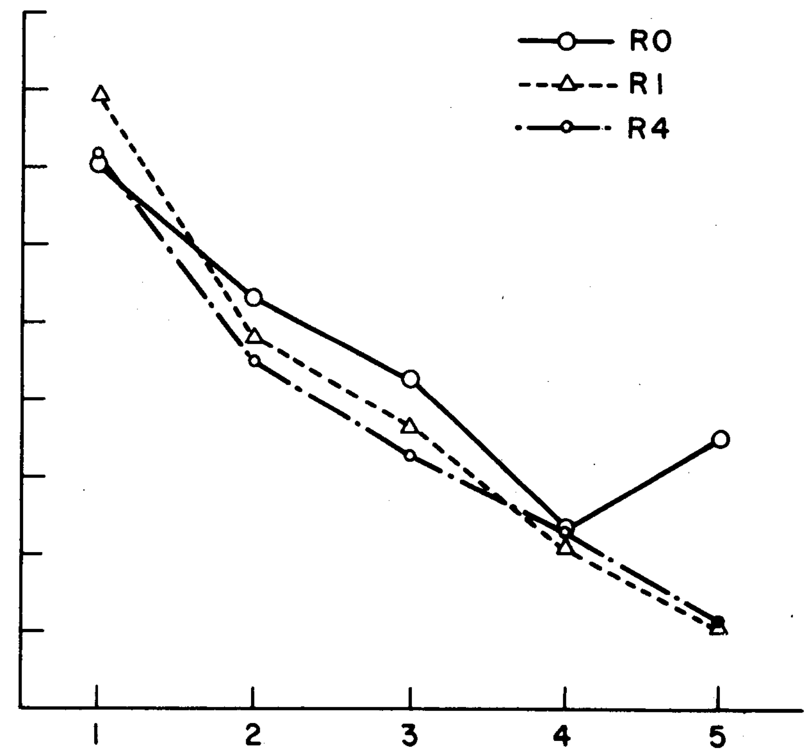

Figure 11b. Mean latency to last letter of response by condition for combined groups

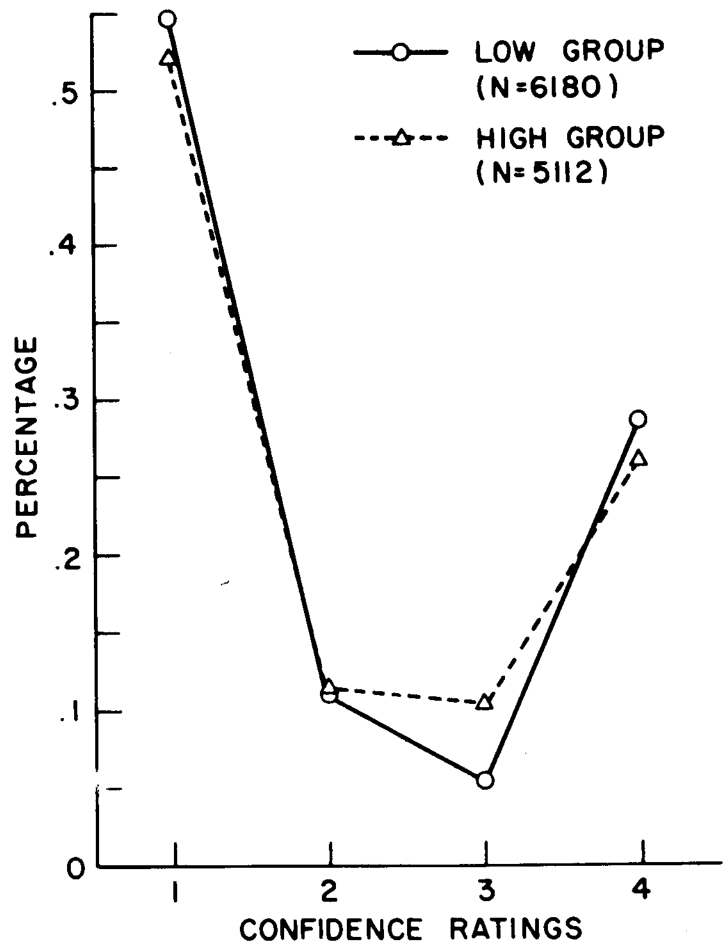

Figure 12. Distribution of confidence ratings for each group

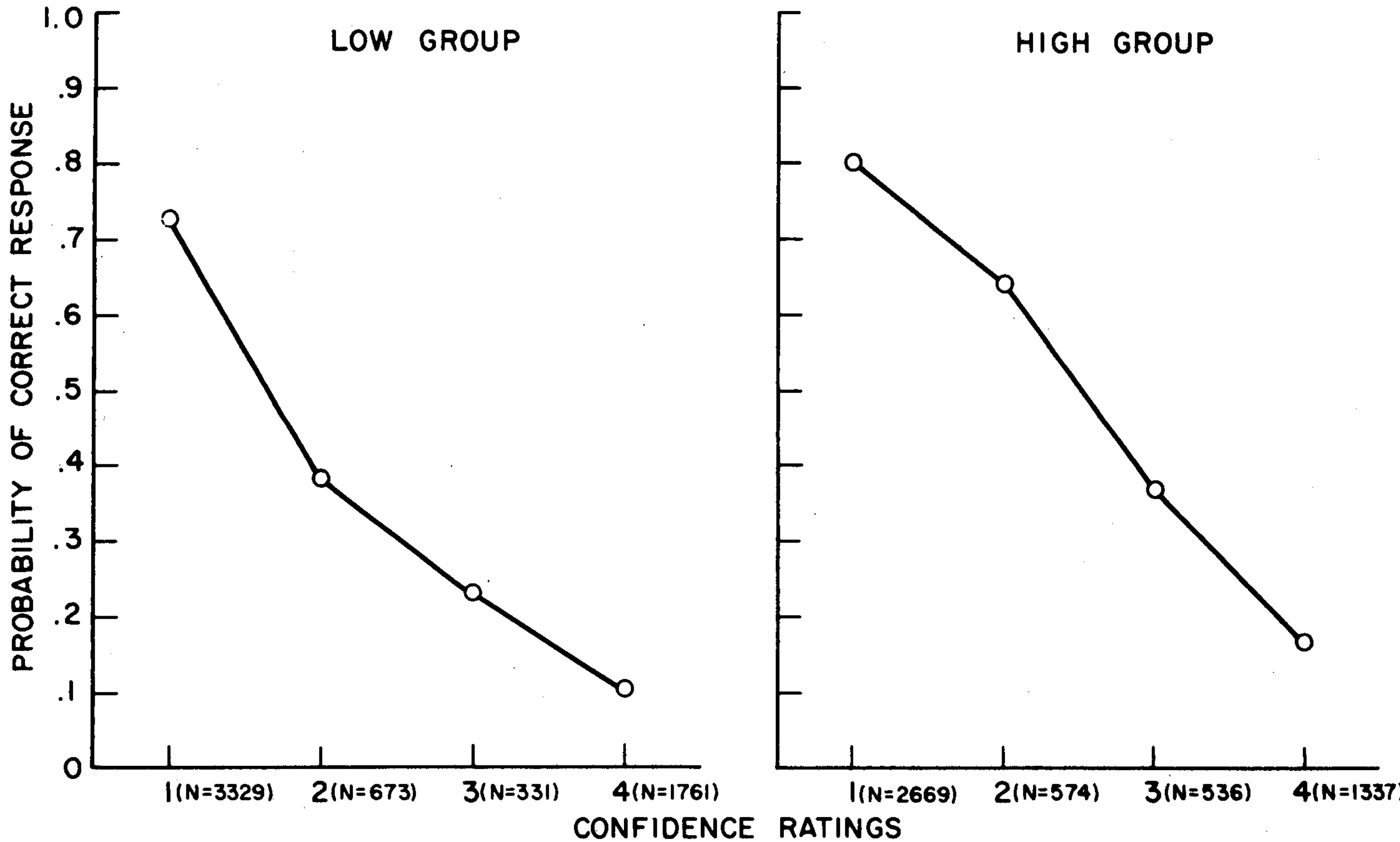

Figure 13. Probability of correct response given the various confidence ratings

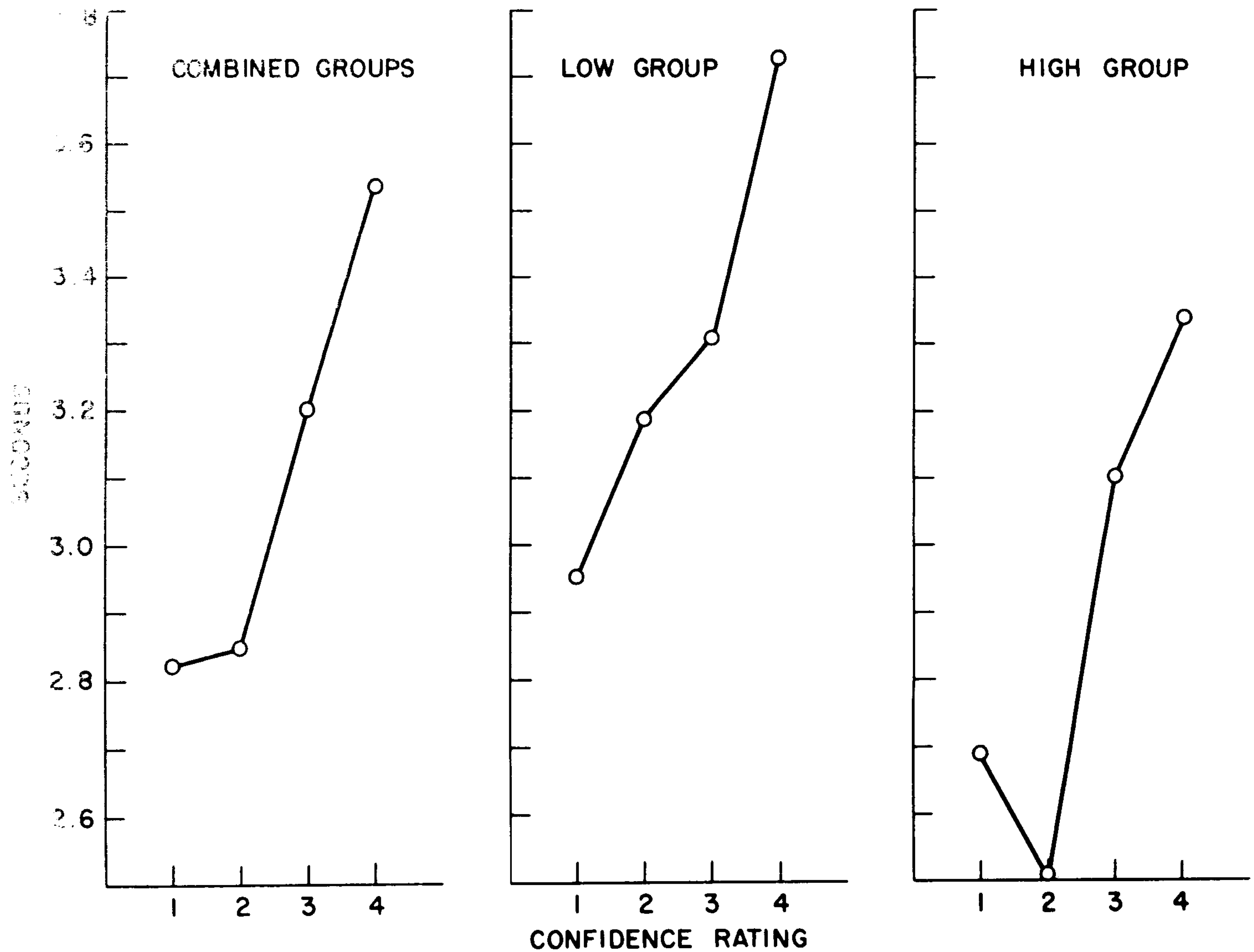

Figure 14. Mean latency to first letter by confidence rating

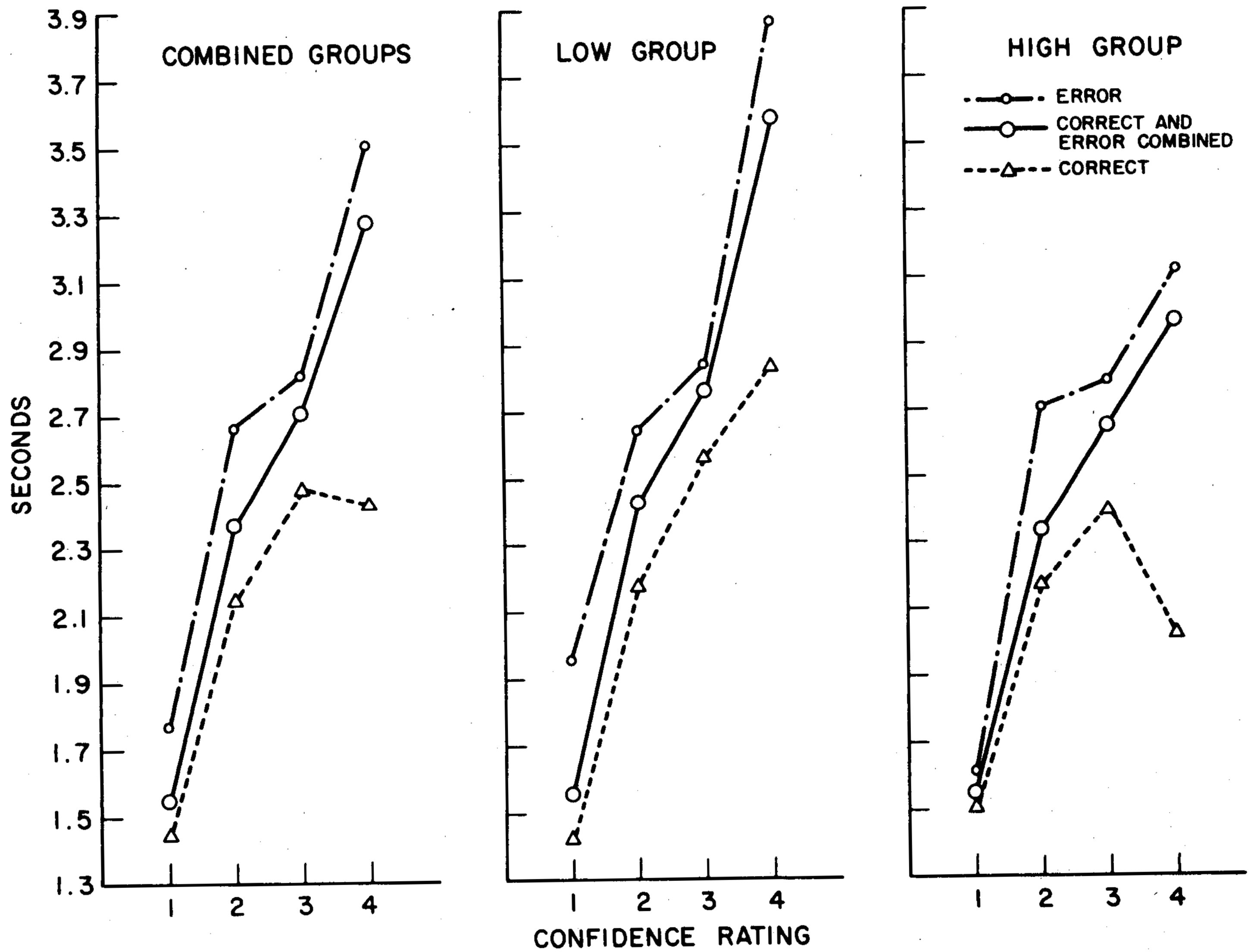

Figure 15. Mean latencies from end of response to confidence rating

CHAPTER IV

DISCUSSION OF EXPERIMENTAL RESULTS

Summary of Major Results

Analyses of variance were computed using the following major dimensions: 3 Conditions (R0, R1, R4); 2 Groups (High and Low); and Weeks (1st through 6th). Using this model, three sets of scores were analyzed: (1) total errors to criterion or five times through each list; (2) total correct on retention tests; (3) difference in number correct on retention test minus the number correct on the first time through a list. In addition, t-tests were run to compare the difference in effects between the various conditions.

Total Errors to Criterion. All three major dimensions were found to be significantly different: Conditions ($F = 20.042$, $p < .001$); Groups ($F = 8.45$, $p < .01$); Weeks ($F = 8.85$, $p < .001$).

T-tests comparing conditions R1 and R4 showed R4 to be significantly superior to R1 for the combined groups, ($t = 3.19$, $p < .001$) and for the Low group ($t = 3.42$, $p < .001$), but not for the High group ($t = .568$).

Retention Test Scores. In this analysis the three major dimensions were also significantly different: Conditions ($F = 3.187$, $p < .05$); Groups ($F = 9.91$, $p < .005$); Weeks ($F = 3.34$, $p < .01$).

T-tests for condition differences were significant beyond the .10 level in only two cases and for only the Low group: R0 vs. R1 ($t = 1.86$, $p < .10$), and R0 vs. R4 ($t = 2.68$, $p < .01$).

Difference Scores. For this analysis the only significant difference was found in the Conditions ($F = 3.51$, $p < .05$). T-tests showed significance in two instances for only the Low group: R0 vs. R1 ($t = 4.04$, $p < .001$), and R0 vs. R4 ($t = 2.31$, $p < .05$).

Group Differences

Our first two major analyses of the data (i.e., errors to criterion and retention test scores) showed substantial differences between the two groups in overall performance and suggested a strong tendency for the condition effects to be more pronounced in the Low group. Both of these results raise interesting problems.

As was mentioned earlier, the attempt to equate the relative difficulty of the words for the two groups was obviously not successful and is reflected in the lower error rate for the High group. In itself, this failure was probably inevitable and not too important. However, this difference in difficulty may be a principal factor in producing the differential effect of the experimental conditions on the groups.

If one examines Figures 4 and 5, which give not only the overall probability of a correct response but also the size of the population contributing to a particular run, the difference is clear. When the N's on these graphs are converted to the actual number of students still running on each day, one finds that on the average by the 3rd Run over half the individuals in the High group had reached criterion, but that only 6 out of 21 had done so in the Low group. By the 4th Run an average of 17 in the High group and only 11 in the Low group had reached criterion. What this means is that a sizeable number of people in the High group are finding the words relatively easy, and that only two or three trials are necessary to produce an almost perfect score. It would seem likely that a possible difference in the potential advantage of one condition over another would be limited here by a "ceiling" effect on learning as a large proportion of students in the High group reach criterion before five runs on a list.

Such an observation does not, of course, rule out the possibility that the conditions actually do affect the two groups in different ways. Further experiments should be able to provide some insight into this problem by further increasing the difficulty level of items for the High group.

Week Differences: Learning-to-Learn

Because of the extended training period preceding the actual start of the experiment, it was not originally expected that the groups would display any significant improvement in performance as they moved through their lists. That this expectation was unfounded seems fairly clear from the analysis and the general downward trend in errors seen in Figure 3--at least for the High group. Although the overall trend is similar for the Low group, the unusual drop between the 3rd and 4th week is rather difficult to explain, except to say that it can probably be attributed to the chance juxtaposition of unusually hard lists in the 3rd week followed by easier ones in week 4.

After puzzling over this apparent learning-to-learn phenomenon and wondering how it could possibly be so potent at such a late stage in training, it was belatedly recalled that the experiment was begun a week after the school's Christmas vacation. In effect, this meant that the students, except for a two-day "warm-up" period just prior to the actual first run in the experiment, had not had any practice for over three weeks. Evidently, this layoff was sufficient to cause them to lose many of the skills necessary for the task.

Condition Effects

The one result which was consistently found through all of the analyses--and which may be clearly seen in Figure 4, for example--is that repetition of error items produced greater learning. This result would seem to contradict Greeno's and Keller's findings that immediate repetition of an item did not produce much learning on the second presentation. It appears likely that the difference in the complexity of response being learned in the present experiment, when compared with that required in Greeno's paired-associate study (i.e., learning the spelling of a word versus learning an arbitrary single digit response) is sufficiently great that the extra immediate practice for spelling was helpful. In Keller's study it should be noted that the experimental procedure perhaps contributed to the apparent ineffectiveness of the immediate repetition of a word. When a child missed a word for the first time in Keller's experiment, he was not given the correct spelling, but was merely told to try again. After the second failure, the correct spelling was presented, and the child was told to copy it. It would seem quite possible that this copying response would involve a minimum of cognitive activity, and hence would perhaps not help the student in his future attempts to recall the word.

In regard to distributed practice on error items, the results of the present experiment are less clear, but seem potentially more interesting. The R4 condition (repetition after four intervening items) was found to be consistently superior to R1 (immediate repetition), but the difference was often not significant. The important question is whether or not this consistent trend represents an actual difference between the effects of R1 and R4. Certainly in the analysis of total errors to criterion and the accompanying t-test comparing R1 and R4 (see Tables 3 and 4), the difference between these conditions for the Low group seems substantial. However, the fact that none of the analyses produced a significant difference between R1 and R4 for the High group indicates that we must be cautious in our inferences. While it seems plausible that the relative ease with which the High group learned the words may have acted to limit the potential real differences between condition effects, we can not be certain that this was actually the case. Therefore in the following discussion our effort to explain the consistent superiority of R4 over R1 admittedly rests on the assumption that this difference in condition effects was not a chance phenomenon.

Any attempt to account for the possible superiority of R4 must of necessity be highly speculative. As a theoretical framework this description will rely strongly on the conceptualization of memory processes proposed by Atkinson and Shiffrin (1967), although no attempt will be made to apply their mathematical models.

If we recall the experimental procedures, we will remember that the subjects were given no indication as to the condition under which a word was presented. They therefore had no idea if a word was going to be repeated--either immediately or a few items later. (This statement is not strictly true after the first run on a list, for there was the possibility that a subject would recall that a particular word was or was not going to be repeated. Interviews with the subjects at the conclusion of the experiment, however, revealed that if such recall was occurring, they were not aware of it.)

It will be recalled that when a subject missed a word, he was given the correct spelling, followed by a six-second study interval before the next item was presented. We shall assume that during this interval whatever strategy the student used to try to learn the word was not dependent on the condition. If one is not willing to accept this assumption for all runs, it certainly is valid for Run 1 on a list when the subject could not have any idea as to which items would be repeated. In any case, we would suggest that the difference in the effects of the R1 and R4 conditions occurs as a result of the difference in the processes used to retrieve the words from memory when they are presented a second time.

As an example, when an R1 word is missed, it is presented immediately again at the end of the study interval. Regardless of what strategies the student may have used to try to commit the word to memory during the study period, an attempt to spell the word at this point would not seem to necessitate any search of long-term memory or any attempt on his part to reconstruct his original strategy. All he needs to be able to do is to keep in his short-term memory, for a period of 10 to 15 seconds, the particular spelling he has just seen. He could seemingly accomplish this simply by engaging in a serial rehearsal of the letters.

On the other hand, the recall of an R4 word on the second presentation during a run is more complex. He has in the interim between the first and second presentation been actively engaged in learning and recalling other words. This activity would seem to preclude any possibility that he may have kept the word in short-term memory through a rehearsal process. More likely, when called upon to spell the word a second time, he must attempt to either retrieve from long-term memory whatever representation of the word he has stored there, or reconstruct the spelling of the word using the strategies he devised earlier. The retrieval process here must be much more similar to that which is ordinarily used in generating the spelling of a word than is the case in the second try on an R1 word.

To summarize briefly, it is being hypothesized that the superiority of the R4 condition perhaps can be accounted for in the kind of practice it provides in <u>retrieving</u> a word from long-term memory; in contrast, the R1 condition does not necessitate such practice since these words can be recalled through a rehearsal process in short-term memory.

The validity of this hypothesis could perhaps be tested if one interspersed a short interference activity after each study interval for the words. This should prevent rehearsal of the words in short-term memory. If one then found that the superiority of R4 over R1 was diminished, a stronger argument could be made that the difference lies in the retrieval processes.

This particular interpretation of the experimental results is quite consistent with the idea stemming from Greeno's study that distributed practice would be superior because it provides training in discrimination. In this case one could speak of the R4 retrieval process as one which constantly necessitates discrimination among similar items in long-term memory. This notion would perhaps have greater validity if we qualified it by asserting that accurate spelling does seem finally to depend on one's ability to make precise discriminations between reasonable alternatives; but this ability to discriminate in turn depends on the efficient and accurate use of memory processes to retrieve these alternatives.

Retention Test Scores

Two important questions need to be discussed in relation to the retention tests, although in neither case are definitive answers readily obvious. The first question is concerned with the decrease on the retention tests of the magnitude of the differences between condition effects; this decrease may be observed in Figures 4 and 5, and in our analysis of test scores. Although there is still a significant overall difference between conditions, the effects are considerably reduced, and there is no longer any significant difference between R1 and R4.

Given the relatively few practice trials that a student had on a word and the length of the retention interval, this decrease is not really surprising. What probably happened is that the subject forgot many of the particular associational and retrieval strategies he had developed and had only partially mastered, and in the retention test was forced to rely on his long-term store of well-learned words and his ability to process words on the basis of the regularity of their spelling.

The second question to be discussed relates to the possible diminution of retention test scores due to the session length. If this does indeed account for the unexpected findings displayed in Figure 6, then it also underscores the need for more research on optimum session length for drills of this type. Perhaps such research would reveal that the length of the session itself is not the most important variable, but rather that the unexpected variation in the required time for a session produces adverse effects. On the other hand, it is also possible that the type of active concentration and attention required in this kind of task is such that brief sessions are best for maximizing the rate of learning.

Latency Data

One of the often-mentioned potential advantages of CAI systems is that machine control of stimulus presentation enables the experimenter to collect latency data heretofore unavailable to the educational researcher. There seems to be little doubt, as Suppes (1964) asserts, that the relatively crude measures of learning we usually use in educational research may often fail to uncover crucial and real differences between, for example, two methods of teaching a mathematical concept.

On the other hand, latency data may reveal little information which is not obtainable from simpler dependent measures. The crucial distinction perhaps lies in the subtlety of the behavioral change the experimenter is trying to detect. Given that one is uncertain as to the magnitude of change one might reasonably expect from the experimental manipulations, the collection of response latency data would seem to be warranted--especially if their collection is relatively inexpensive and simple as it is in CAI. There is obviously no guarantee that such data will be any more informative than other measures.

In this experiment it seems fairly clear that the latency data does not reveal important differences which were not detected by the correct/error measure. As was pointed out earlier, this failure may in part be due to the way the first latency was measured; or it could simply be due to the fact that there were not any real significant differences in processing time required for the various conditions.

Nevertheless, the fact that the measured latencies in this experiment didn't seem to disclose unexpected information, or that they were primarily a means of confirming information gained from simpler dependent measures, does not negate their potential value in CAI research.

Confidence Rating Measure

As was expected for reasons pointed out earlier, it is difficult to interpret the results for this particular dependent measure. If one looks simply at the overall summary statistics such as those found in Figure 13, it appears that the confidence ratings and the probability of a correct response are related in an orderly, linear fashion. However, the wide variation in individual accuracy and the low usage of categories two and three would seem to place severe limitations on the usefulness of this particular measure--at least as it was obtained in this experiment.

If a confidence rating measure is to be a valuable dependent variable, then it would seem desirable that steps be taken to increase individual accuracy and to encourage wider use of more than two categories. Perhaps this objective could be accomplished through some sort of feedback to the individual as to the accuracy of his confidence rating. (For example, see Phillips, Shiffrin and Atkinson, 1967). In our experimental situation there was no particular incentive for an individual to try to maximize the accuracy of his rating or to use all of the categories, since the evaluation of his response was based strictly on his spelling. But it is also possible that for students of this age group and ability that the degree of their uncertainty or confidence is bipolar, and that they find it difficult to make finer discriminations.

From a pedagogical point of view there can be little doubt that children should be encouraged to assess the accuracy of their spelling of a word, and it seems likely that this decision process could be improved through proper training. Further experimentation in which such training was a principal variable could possibly provide valuable information.

It is also possible that other indirect measures than those employed would yield more precise knowledge as to the state of learning in the individual. For example, it would be a relatively simple matter to allow the student to start over when he felt he had made a mistake on a word. Some combination of measures of the time he took on the word and the number of re-starts he used might be highly informative.

In conclusion, it should be emphasized that the problem of trying to obtain more sensitive measures of learning is very important--particularly to CAI. The power of the computer to make decisions on optimum selection and sequencing of materials will only be realized when, and if, we are able to find the proper dependent measures which adequately characterize the individual's current state of learning.

Appendix A

The number after each of the words is the percentage of students on that grade level who spelled the word correctly.

Source: New Iowa Spelling Scale

4th Grade Level

List 1.

1.	burned	40
2.	louder	40
3.	truth	41
4.	dresses	42
5.	score	43
6.	young	44
7.	lend	45
8.	tore	46
9	camel	47
10.	starting	48
11.	fresh	49
12	month	50

List 2.

1.	careless	40
2.	perhaps	40
3.	wear	41
4.	follow	42
5.	tables	43
6.	church	44
7.	main	45
8.	trust	46
9.	began	47
10.	thinking	48
11.	driving	49
12.	higher	50

List 3.

1.	climbing	40
2.	present	40
3.	windy	41
4.	heard	42
5.	tired	43
6.	carpet	44
7.	places	45
8.	uncle	46
9.	blind	47
10.	tune	48
11.	drinking	49
12.	inch	50

List 4.

1.	clothing	40
2.	shirt	40
3.	arithmetic	41
4.	iron	42
5.	together	43
6.	escape	44
7.	stairs	45
8.	wheel	46
9.	bought	47
10.	copy	48
11.	blame	49
12	rule	50

List 5.

1.	fifty	40
2.	printing	40
3.	coin	41
4.	reward	42
5.	until	43
6.	everywhere	44
7.	stuck	45
8.	sunny	46
9.	color	47
10.	bottle	48
11.	bigger	49
12.	joke	50

List 6.

1.	join	40
2.	speak	40
3.	cotton	41
4.	proud	42
5.	twenty	43
6.	leaving	44
7.	tiny	45
8.	writing	46
9.	space	47
10.	yourself	48
11.	basket	49
12.	queen	50

Appendix A (cont.)

5th Grade Level

List 7.			List 8.			List 9.		
1.	puzzle	40	1.	harbor	40	1.	fields	40
2.	complain	41	2.	theater	41	2.	swiftly	41
3.	factory	42	3.	curve	42	3.	double	42
4.	human	42	4.	insist	42	4.	laughed	42
5.	seldom	43	5.	vanish	43	5.	contains	43
6.	rare	44	6.	peanut	44	6.	protect	44
7.	earned	45	7.	coward	45	7.	climbing	45
8.	towel	46	8.	vacation	46	8.	animal	46
9.	quit	47	9.	sailor	47	9.	using	47
10.	circle	48	10.	sign	48	10.	either	48
11.	knock	49	11.	rainy	49	11.	since	49
12.	ladies	50	12.	level	50	12.	tried	50

List 10.			List 11.			List 12.		
1.	cottage	40	1.	control	40	1.	admit	40
2.	selfish	41	2.	really	41	2.	questions	41
3.	figure	42	3.	foolish	42	3.	habit	42
4.	machine	42	4.	partner	42	4.	rough	42
5.	dining	43	5.	lying	43	5.	oranges	43
6.	telephone	44	6.	wilderness	44	6.	agreed	44
7.	picture	45	7.	quarter	45	7.	shoulder	45
8.	cloudy	46	8.	explain	46	8.	pencils	46
9.	whose	47	9.	elect	47	9.	nicest	47
10.	flight	48	10.	kitchen	48	10.	velvet	48
11.	chief	49	11.	howling	49	11.	interest	49
12.	steel	50	12.	taught	50	12.	thrill	50

Appendix A (cont.)

6th Grade Level

	List 13.			List 14.			List 15.	
1.	appointment	40	1.	astonish	40	1.	barely	40
2.	families	41	2.	fifteenth	41	2.	further	41
3.	operate	42	3.	planning	42	3.	prompt	42
4.	cruel	43	4.	doubt	43	4.	guilty	43
5.	release	44	5.	rescue	44	5.	reflection	44
6.	independent	45	6.	believe	45	6.	correction	45
7.	breeze	46	7.	carrying	46	7.	condition	46
8.	elevator	47	8.	furniture	47	8.	dentist	47
9.	natural	48	9.	pledge	48	9.	position	48
10.	review	49	10.	length	49	10.	addition	49
11.	depth	50	11.	decorate	50	11.	direction	50
12.	importance	50	12.	increase	50	12.	junior	50

	List 16.			List 17.			List 18.	
1.	cabinet	40	1.	continued	40	1.	exact	40
2.	include	41	2.	injure	41	2.	journey	41
3.	accept	42	3.	convince	42	3.	creature	42
4.	lettuce	43	4.	private	43	4.	fortune	43
5.	practice	44	5.	neither	44	5.	memory	44
6.	janitor	45	6.	nickel	45	6.	meant	45
7.	discovery	46	7.	avoid	46	7.	attention	46
8.	chimney	47	8.	easily	47	8.	blossom	47
9.	entertain	48	9.	recess	48	9.	search	48
10.	rotten	49	10.	national	49	10.	capital	49
11.	division	50	11.	energy	50	11.	expecting	50
12.	loan	50	12.	musical	50	12.	potato	50

REFERENCES

Atkinson, R. C. and Hansen, D. N. Computer-assisted instruction in initial reading: the Stanford project. Reading Research Quarterly, 1966, 2, 5-25.

Atkinson, R. C. and Shiffrin, R. M. Human memory: a proposed system and its control processes. Technical Report No. 110, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1967.

Beltramo, Louise. An Alphabetic Approach to the Teaching of Reading in Grade One. Unpublished Doctor's Thesis, State University of Iowa, 1954.

Bernbach, H. Decision processes in memory. Psychological Review, 1967, in press.

Coulson, J. E. (Ed.) Programmed Learning and Computer-Based Instruction. New York: John Wiley & Sons, Inc., 1962.

Fishman, Elizabeth. Massed vs. distributed practice in computerized spelling drills. Unpublished manuscript, 1967.

Gates, A. I. Generalization and Transfer in Spelling. New York: Teachers College, 1935.

Glaser, R. (Ed.) Teaching Machines and Programmed Learning, II. Washington, D. C.: Department of Audiovisual Instruction, National Education Association of the United States, 1965.

Greene, H. A. The New Iowa Spelling Scale. Iowa City, Iowa: State University of Iowa, 1954.

Greeno, J. G. Paired-associate learning with massed and distributed repetition of items. Journal of Experimental Psychology, 1964, 67, 286-295.

Hahn, W. P. Phonics: a boon to spelling? Elementary School Journal, 1964, 64, 383-386.

Hanna, P. R., Hanna, Jean S., Hodges, R. E., and Rudorf, E. H. Phoneme-Grapheme Correspondences as Cues to Spelling Improvement. Washington: U. S. Department of Health, Education, and Welfare, Office of Education, 1966.

Horn, E. "Spelling" in Encyclopedia of Educational Research. New York: The Macmillan Company, 1960, 1337-1354.

Keller, L. Two correction procedures in spelling drills using a CAI system. Unpublished manuscript, 1966.

McNemar, Quinn. Psychological Statistics. New York: John Wiley & Sons, 1962.

Phillips, J. L., Shiffrin, R. M., and Atkinson, R. C. The effects of list length on short-term memory. Journal of Verbal Learning and Verbal Behavior, 1967, in press.

Sartorius, Ina C. Generalization in Spelling. New York: Teachers College, 1931.

Suppes, P. Modern learning theory and the elementary-school curriculum. American Educational Research Journal, 1964, 1, 79-83.

Suppes, P. The uses of computers in education. Scientific American, 1966, 215, 207-220.

Suppes, P. The computer and excellence. Saturday Review, January 14, 1967, 46-50.

Suppes, P., Hyman, L., Jerman, M. Linear structural models for response and latency performance in arithmetic. Technical Report No. 100, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1966.

Suppes, P., Jerman, M., and Groen, G. Arithmetic drills and review on a computer-based teletype. Technical Report No. 83, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1965.

Wilson, H. A., and Atkinson, R. C. Computer-based instruction in in initial reading: a progress report of the Stanford project. In H. Levin and Joanna Williams (Eds.), Basic Studies on Reading, 1967, in press.

(Continued from inside front cover)

36 G. H. Bower. Response strengths and choice probability: A consideration of two combination rules. December 19, 1960. (In E. Nagel, P. Suppes, and A. Tarski (Eds.), Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress. Stanford Univ. Press, 1962. Pp. 400-412)

37 G. H. Bower. Application of the all-or-none conditioning model to the learning of compound responses. June 7, 1961.

38 P. Suppes and M. Schlag-Rey. Test of some learning models for double contingent reinforcement. August 15, 1961. (Psychol. Rep., 1962,

39 P. Suppes and R. Ginsberg. A fundamental property of all-or-none models, binomial distribution of responses prior to conditioning, with application to concept formation in children. September 20, 1961. (Psychol. Rev., 1963, 70, 139-161)

40 J. Theios. A three-state Markov model for learning. September 22, 1961. (Simple conditioning as two-stage all-or-none learning, Psychol. Rev., 1963, 70, 403-417)

41 G. H. Bower. General three-state Markov learning models. September 26, 1961.

42 R. C. Atkinson. A variable threshold model for signal detection. November 17, 1961.

43 R. C. Atkinson. Mathematical models in research on perception and learning. December 25, 1961. (In M. H. Marx (Ed.), Theories in Contemporary Psychology. New York: Macmillan Co., 1963. Pp. 551-564)

44 P. Suppes. Towards a behavioral foundation of mathematical proofs. January 2, 1962. (In K. Ajdukiewicz (Ed.), The Foundations of Statements and Decisions: Proceedings of the International Colloquium on Methodology of Sciences, September 18-23, 1961. Warszawa: PWN-Polish Scientific Publishers, 1965. Pp. 327-341)

45 P. Suppes and J. L. Zinnes. Basic measurement theory. March 15, 1962. (Chapter I in R. R. Bush, E. H. Galanter, and R. D. Luce (Eds.), Handbook of Mathematical Psychology, Vol. I. New York: John Wiley, 1963)

46 R. C. Atkinson, E. C. Carterette, and R. A. Kinchla. Sequential phenomena in psychophysical judgments: a theoretical analysis. April 20, 1962. (Institute of Radio Engineers Transactions on Information Theory, 1962, IT-8, S 155-162)

47 R. C. Atkinson. A variable sensitivity theory of signal detection. May 18, 1962. (Psychol. Rev., 1963, 70, 91-106)

48 R. C. Atkinson and W. K. Estes. Stimulus sampling theory. July 1, 1962. (Chapter 10 in R. R. Bush, G. H. Galanter, and R. D. Luce (Eds.), Handbook of Mathematical Psychology, Vol. II. New York: Wiley, 1963)

49 P. Suppes, E. Crothers, R. Weir, and E. Trager. Some quantitative studies of Russian consonant phoneme discrimination. September 14, 1962.

50 R. C. Atkinson and R. C. Calfee. Mathematical learning theory. January 2, 1963. (In B. B. Wolman (Ed.), Scientific Psychology. New York: Basic Books, Inc., 1965. Pp. 254-275)

51 P. Suppes, E. Crothers, and R. Weir. Application of mathematical learning theory and linguistic analysis to vowel phoneme matching in Russian words. December 28, 1962.

52 R. C. Atkinson, R. Calfee, G. Sommer, W. Jeffrey and R. Shoemaker. A test of three models for stimulus compounding with children. January 29, 1963. (J. exp. Psychol., 1964, 67, 52-58)

53 E. Crothers. General Markov models for learning with inter-trial forgetting. April 8, 1963.

54 J. L. Myers and R. C. Atkinson. Choice behavior and reward structure. May 24, 1963. (Journal math. Psychol., 1964, 1, 170-203)

55 R. E. Robinson. A set-theoretical approach to empirical meaningfulness of measurement statements. June 10, 1963.

56 E. Crothers, R. Weir and P. Palmer. The role of transcription in the learning of the orthographic representations of Russian sounds. June 17, 1963.

57 P. Suppes. Problems of optimization in learning a list of simple items. July 22, 1963. (In Maynard W. Shelly, II and Glenn L. Bryan (Eds.), Human Judgments and Optimality. New York: Wiley. 1964. Pp. 116-126)

58 R. C. Atkinson and E. J. Crothers. Theoretical note: all-or-none learning and intertrial forgetting. July 24, 1963.

59 R. C. Calfee. Long-term behavior of rats under probabilistic reinforcement schedules. October 1, 1963.

60 R. C. Atkinson and E. J. Crothers. Tests of acquisition and retention, axioms for paired-associate learning. October 25, 1963. (A comparison of paired-associate learning models having different acquisition and retention axioms, J. math. Psychol., 1964, 1, 285-315)

61 W. J. McGill and J. Gibbon. The general-gamma distribution and reaction times. November 20, 1963. (J. math. Psychol., 1965, 2, 1-18)

62 M. F. Norman. Incremental learning on random trials. December 9, 1963. (J. math. Psychol., 1964, 1, 336-351)

63 P. Suppes. The development of mathematical concepts in children. February 25, 1964. (On the behavioral foundations of mathematical concepts. Monographs of the Society for Research in Child Development, 1965, 30, 60-96)

64 P. Suppes. Mathematical concept formation in children. April 10, 1964. (Amer. Psychologist, 1966, 21, 139-150)

65 R. C. Calfee, R. C. Atkinson, and T. Shelton, Jr. Mathematical models for verbal learning. August 21, 1964. (In N. Wiener and J. P. Schoda (Eds.), Cybernetics of the Nervous System: Progress in Brain Research. Amsterdam, The Netherlands: Elsevier Publishing Co., 1965. Pp. 333-349)

66 L. Keller, M. Cole, C. J. Burke, and W. K. Estes. Paired associate learning with differential rewards. August 20, 1964. (Reward and information values of trial outcomes in paired associate learning. (Psychol. Monogr., 1965, 79, 1-21)

67 M. F. Norman. A probabilistic model for free-responding. December 14, 1964.

68 W. K. Estes and H. A. Taylor. Visual detection in relation to display size and redundancy of critical elements. January 25, 1965, Revised 7-1-65. (Perception and Psychophysics, 1966, 1, 9-16)

69 P. Suppes and J. Donio. Foundations of stimulus-sampling theory for continuous-time processes. February 9, 1965.

70 R. C. Atkinson and R. A. Kinchla. A learning model for forced-choice detection experiments. February 10, 1965. (Br. J. math stat. Psychol., 1965, 18, 184-206)

71 E. J. Crothers. Presentation orders for items from different categories. March 10, 1965.

72 P. Suppes, G. Groen, and M. Schlag-Rey. Some models for response latency in paired-associates learning. May 5, 1965. (J. math. Psychol., 1966, 3, 99-128)

73 M. V. Levine. The generalization function in the probability learning experiment. June 3, 1965.

74 D. Hansen and T. S. Rodgers. An exploration of psycholinguistic units in initial reading. July 6, 1965.

75 B. C. Arnold. A correlated urn-scheme for a continuum of responses. July 20, 1965.

76 C. Izawa and W. K. Estes. Reinforcement-test sequences in paired-associate learning. August 1, 1965.

77 S. L. Blehart. Pattern discrimination learning with Rhesus monkeys. September 1, 1965.

78 J. L. Phillips and R. C. Atkinson. The effects of display size on short-term memory. August 31, 1965.

79 R. C. Atkinson and R. M. Shiffrin. Mathematical models for memory and learning. September 20, 1965.

80 P. Suppes. The psychological foundations of mathematics. October 25, 1965.

(Continued on back cover)

(Continued from inside back cover)

81 P. Suppes. Computer-assisted instruction in the schools: potentialities, problems, prospects. October 29, 1965.

82 R. A. Kinchla, J. Townsend, J. Yellott, Jr., and R. C. Atkinson. Influence of correlated visual cues on auditory signal detection. November 2, 1965. (Perception and Psychophysics, 1966, 1, 67-73)

83 P. Suppes, M. Jerman, and G. Groen. Arithmetic drills and review on a computer-based teletype. November 5, 1965.

84 P. Suppes and L. Hyman. Concept learning with non-verbal geometrical stimuli. November 15, 1965.

85 P. Holland. A variation on the minimum chi-square test. November 18, 1965.

86 P. Suppes. Accelerated program in elementary-school mathematics -- the second year. November 22, 1965.

87 P. Lorenzen and F. Binford. Logic as a dialogical game. November 29, 1965.

88 L. Keller, W. J. Thomson, J. R. Tweedy, and R. C. Atkinson. The effects of reinforcement interval on the acquisition of paired-associate responses. December 10, 1965.

89 J. I. Yellott, Jr. Some effects on noncontingent success in human probability learning. December 15, 1965.

90 P. Suppes and G. Groen. Some counting models for first-grade performance data on simple addition facts. January 14, 1966.

91 P. Suppes. Information processing and choice behavior. January 31, 1966.

92 G. Groen and R. C. Atkinson. Models for optimizing the learning process. February 11, 1966.

93 R. C. Atkinson and D. Hansen. Computer-assisted instruction in initial reading: Stanford project. March 17, 1966.

94 P. Suppes. Probabilistic inference and the concept of total evidence. March 23, 1966.

95 P. Suppes. The axiomatic method in high-school mathematics. April 12, 1966.

96 R. C. Atkinson, J. W. Breisford, and R. M. Shiffrin. Multi-process models for memory with applications to a continuous presentation task. April 13, 1966.

97 P. Suppes and E. Crothers. Some remarks on stimulus-response theories of language learning. June 12, 1966.

98 R. Bjork. All-or-none subprocesses in the learning of complex sequences. June 22, 1966.

99 E. Gammon. The statistical determination of linguistic units. July 1, 1966.

100 P. Suppes, L. Hyman, and M. Jerman. Linear structural models for response and latency performance in arithmetic. July 29, 1966.

101 J. L. Young. Effects of intervals between reinforcements and test trials in paired-associate learning. August 1, 1966.

102 H. A. Wilson. An investigation of linguistic unit size in memory processes. August 3, 1966.

103 J. T. Townsend. Choice behavior in a cued-recognition task. August 8, 1966.

104 W. H. Batchelder. A mathematical analysis of multi-level verbal learning. August 9, 1966.

105 H. A. Taylor. The observing response in a cued psychophysical task. August 10, 1966.

106 R. A. Bjork. Learning and short-term retention of paired associates in relation to specific sequences of interpresentation intervals. August 11, 1966.

107 R. C. Atkinson and R. M. Shiffrin. Some Two-process models for memory. September 30, 1966.

108 P. Suppes and C. Ihrke. Accelerated program in elementary-school mathematics--the third year. January 30, 1967.

109 P. Suppes and I. Rosenthal-Hill. Concept formation by kindergarten children in a card-sorting task. February 27, 1967.

110 R. C. Atkinson and R. M. Shiffrin. Human memory: a proposed system and its control processes. March 21, 1967

111 Theodore S. Rodgers. Linguistic considerations in the design of the Stanford computer-based curriculum in initial reading. June 1, 1967

112 Jack M. Knutson. Spelling drills using a computer-assisted instructional system. June 30, 1967.