DOCUMENT RESUME

ED 187 746 TM 800 248

AUTHOR Roid, Gale H.; And Others
TITLE Algorithms for Developing Test Questions from Sentences in Instructional Materials: An Extension of an Earlier Study.
INSTITUTION Oregon State System of Higher Education, Monmouth.
SPONS AGENCY Advanced Research Projects Agency (DOD), Washington, D.C.
REPORT NO NPRDC-TR-80-11
PUB DATE Jan 80
CONTRACT MDA-903-77-C-0189
NOTE 28p.; For related document, see ED 163 024.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Algorithms; Computer Assisted Testing; *Criterion Referenced Tests; Difficulty Level; Form Classes (Languages); Higher Education; High Schools; *Multiple Choice Tests; *Test Construction; *Test Items

ABSTRACT

An earlier study was extended and replicated to examine the feasibility of generating multiple-choice test questions by transforming sentences from prose instructional material. In the first study, a computer-based algorithm was used to analyze prose subject matter and to identify high-information words. Sentences containing selected words were then transformed into multiple-choice items by four writers who generated foils or question alternatives informally and by an algorithmic method. These items were then organized into tests and administered to 24 college students before and after they had studied the instructional materials. In this replication, the tests were administered to 249 high school students, and results were combined with those obtained earlier. This provided stable estimates of item difficulty. Results supported those obtained earlier. Thus, it appears that this item-writing technique is feasible and that algorithmic methods of generating foils produce items of reasonably good quality. (The prose passage used in the study and examples of test items are appended). (Author/CTM)

NPRDC Technical Report 80-11 January 1980

ALGORITHMS FOR DEVELOPING TEST QUESTIONS FROM SENTENCES IN INSTRUCTIONAL MATERIALS: AN EXTENSION OF AN EARLIER STUDY

Gale Roid
Tom Haladyna
Oregon State System of Higher Education
Monmouth, Oregon 97361

Patrick Finn
State University of New York at Buffalo
Buffalo, New York 14260

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Navy Personnel Research and Development Center under Contract MDA-903-77-C-0189.

Reviewed by
John D. Ford, Jr.

Approved by
James J. Regan
Technical Director

Navy Personnel Research and Development Center
San Diego, California 92152

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1 REPORT NUMBER
NPRDC TR 80-11

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4 TITLE (and Subtitle)
ALGORITHMS FOR DEVELOPING TEST QUESTIONS FROM SENTENCES IN INSTRUCTIONAL MATERIALS: AN EXTENSION OF AN EARLIER STUDY

5. TYPE OF REPORT & PERIOD COVERED
Sep 1977-Sep 78

6. PERFORMING ORG. REPORT NUMBER

7 AUTHOR(s)
Gale H. Roid
Tom Haladyna
Patrick Finn

8. CONTRACT OR GRANT NUMBER(s)
MDA-903-77-C-0189

9 PERFORMING ORGANIZATION NAME AND ADDRESS
Oregon State System of Higher Education
Monmouth, Oregon 97361

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
62709N-RPA.3354

11 CONTROLLING OFFICE NAME AND ADDRESS
Defense Advanced Research Projects Agency
Arlington, Virginia 22209

12. REPORT DATE
January 1980

13. NUMBER OF PAGES
24

14 MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
Navy Personnel Research and Development Center
San Diego, California 92152

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16 DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17 DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18 SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)
Criterion-referenced tests
Item-writing methods
Automated algorithms for writing items
Item-objective congruence
Testing prose material
Multiple-choice test items

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)
The purpose of this effort was to extend or replicate an earlier study that examined the feasibility of generating multiple-choice test questions by transforming sentences from prose instructional materials. In that study, a computer-based algorithm was used to analyze prose subject matter and to identify high-information words. Sentences containing selected words were then transformed into multiple-choice items by four writers who generated foils or question alternatives informally

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

FEB 13 1980

and by an algorithmic method. These items were then organized into tests and administered to 24 college students before and after they had studied the instructional materials.

In this replication, the tests were administered to 249 high school students, and results were combined with those obtained earlier. This provided stable estimates of item difficulty. Results supported those obtained earlier. Thus, it appears that this item-writing technique is feasible and that algorithmic methods of generating foils produce items of reasonably good quality.

FOREWORD

This research and development was conducted under the sponsorship of the Defense Advanced Research Projects Agency and is related to studies of criterion-referenced testing being conducted at this Center. Information resulting from this testing will be incorporated in a testing manual being prepared by the Navy Personnel Research and Development Center. This manual will be used operationally by the Chief of Naval Education and Training, the Chief of Naval Technical Training, and the Chief of Naval Education and Training Support (specifically, the Instructional Program Development Centers).

A previous report, NPRDC TR 78-23 of June 1978, described the beginning phases of a contractual effort aimed at examining the qualities of test questions written from a variety of methods. This report describes a replication and extension of that work. Results will be considered in further development of algorithmic procedures for generating test questions from prose materials.

Appreciation is expressed to Dr. John R. Bormuth of the University of Chicago, and Dr. Jason Millman of Cornell University, who were consultants for this project.

Dr. Pat-Anthony Federico of this Center served as the Contracting Officer Technical Representative.

DONALD F. PARKER
Commanding Officer

SUMMARY

Problem and Background

Methods for writing test questions or items, particularly for criterion-referenced testing, are needed that are (1) based on a logically defined relationship between the instructional materials and the test items written to assess learning from those materials, and (2) capable of producing items that can be easily replicated by many test developers. Such methods should allow tests to become more scientific instruments and contribute to the advancement of instructional research, educational evaluation, and the use of test data in forming public policy.

In an earlier study (NPRDC TR 78-23), an attempt was made to refine a method of objectively generating multiple-choice test questions by transforming sentences from prose instructional materials and developing foils or question alternatives by an algorithmic method. In that study, selected instructional material was computer-analyzed to identify high information words--those that are relatively rare in American English--and to determine the text frequency of those words. Twenty high information nouns and adjectives--10 rare singletons and 10 keywords--were selected for use as questions words. Singletons are high information words that occur only once in a passage; and keywords, those that occur more than once. Twenty sentences were then selected for transformation into items by four item writers. Five of these sentences included rare singleton nouns; five, rare singleton adjectives; five, keyword nouns; and five, keyword adjectives.

The four item writers transformed the selected sentences by substituting the question words with wh-words (who, what, etc.), and generated item foils or response alternatives both informally and with an algorithmic method. This resulted in 160 items--20 selected sentences transformed by four item writers using two foil methods--that were organized into eight 20-item test forms. These test forms were administered to 24 subjects--three to each form--before (pretest) and after (posttest) they studied the instructional material. Care was taken to ensure that students completed different test forms on the two test occasions. Average pretest and posttest item difficulty, as determined by the percentage of subjects who answered the question correctly, were computed for items (1) produced by each of the four writers, (2) derived from each of the four types of question words, and (3) with foils generated by each of the two methods.

Results indicated that rare singleton nouns and adjectives and keyword adjectives are promising candidates for use as question words in developing questions that test learning from prose. Keyword nouns, however, are not good candidates. It was concluded that the methods used to generate foils algorithmically were feasible. Although foils produced by these methods were somewhat easier than those generated by item writers, they still appeared to produce a significant shift in difficulty from pretest to posttest when instruction was provided between testing sessions.

Purpose

The purpose of this study was to extend or replicate the earlier study. It is expected that the results will form the basis for additional development of algorithmic procedures for generating test questions from prose materials.

Approach

The eight forms were administered to 249 high school students before and after they had studied the instructional material. For both pre- and posttest, about 30 students were randomly assigned to each of the test forms. Care was taken, however, to ensure that the forms administered to each subject on the two test occasions were different.

To obtain stable estimates of item difficulty, test results from the earlier study were combined with those obtained in this study. Thus, the total number of subjects was 273 (24 college students and 249 high school students). A repeated measures analysis of variance was used to examine differences in item difficulties between (1) the four item writers, (2) the two parts of speech of question words, (3) the two types of text frequencies (keyword and rare singletons), (4) the two foil types, and (5) the two test occasions.

Results

1. Items based on rare singleton nouns and adjectives and keyword adjectives showed a significant change in item difficulty from pretest to posttest, indicating that such items are useful in learning from the type of prose used in the study.

2. Items derived from keyword nouns produced low quality items, primarily because the sentences they occurred in were usually introductory sentences of a general nature.

3. The two types of foils proved to be almost equally effective for learning, as evidenced by the similarity in posttest item difficulty. Those generated by item writers, however, were considerably harder on the pretest and showed a higher change in item difficulty from pretest to posttest than did those generated algorithmically.

4. No significant differences between item writers were found, indicating that the sentence transformation methods employed apparently neutralized the effects of item writer bias that has been found in other studies of item writing.

Conclusions

The concept of using a computer-based algorithm to analyze prose instructional materials and to identify high information words appears to be workable. High information rare singleton nouns or adjectives, as well as keyword adjectives that occur no more than three times, appear to be good candidates for question words. Keyword nouns, however, apparently are not good candidates, particularly when they occur in general introductory sentences.

Recommendations

1. Rare singleton nouns and adjectives and keyword adjectives that occur infrequently in instructional material should be used to select sentences from prose passages for transformation into questions that measure reading comprehension. Keyword nouns should not be used, particularly when they occur in general introductory sentences.

2. Methods of algorithmically generating foils for multiple-choice versions of sentence-derived questions should be further refined and applied in a variety of subject matter areas.

CONTENTS

LIST OF TABLES

INTRODUCTION

Problem

Methods for writing test questions or items, particularly for criterion-referenced testing, are needed that are (1) based on a logically defined relationship between the instructional materials and the test items written to assess learning from those materials, (2) defined by a set of operations open to public inspection, and (3) capable of producing items that can be easily replicated by many test developers. Such methods should allow tests to become more scientific instruments and contribute to the advancement of instructional research, educational evaluation, and the use of test data in forming public policy.

Background

Roid and Finn (1978) attempted to refine a method of objectively generating multiple-choice test questions by transforming sentences from prose instructional materials and developing foils or question alternatives by an algorithmic method. A prose passage on insect development (see appendix), which was written for approximately the high school level, was selected for use in the Roid and Finn study. Items (stems and foils) to test learning from this passage were developed using the following procedure:

1. The selected material was computer-analyzed to identify high information words--those that are relatively rare in American English--and to determine the text frequency of those words. Twenty high information nouns and adjectives--10 rare singletons and 10 keywords--were selected for use as question words. Singletons are high information words that occur only once in a passage; and keywords, words that occur more than once.

2. Twenty sentences were then selected for transformation into multiple-choice items by four item writers. Five of these sentences included rare singleton nouns; five, rare singleton adjectives; five, keyword nouns; and five, keyword adjectives.

3. The stems for these multiple-choice items were produced by substituting the question words with wh-words (who, what, etc.). For example, the rare singleton "silverfish" appeared in the following sentence: "The most primitive insects, such as the silverfish, do not go through metamorphosis." For this sentence, one writer produced the following item stem: "The most primitive insects, such as what, do not go through metamorphosis?" Next, for each of the 20 stem items produced, each writer produced two sets of foils or alternatives. One set was produced informally by the writer; and the other, by an algorithmic method. For example, for the above item stem, the writer/author produced the following foils:

   a. Informally--Butterflies, Silverfish, Canine, and Cicadas.
   b. Algorithmically--Silverfish, Females, Individuals, and Wasps.

This process resulted in 160 multiple-choice items: 20 selected sentences transformed by four item writers using two foil methods. For a given instance, the stems, as well as the foils produced informally by the writers, were comparable but not identical. The foils produced algorithmically, however, were the same across items/writers. Examples are provided in the appendix.

To generate foils for the rare singleton and keyword nouns, those selected as question words were classified semantically using the method developed by Fredericksen (1975), which is shown in Figure 1. To illustrate, using this method, the singleton noun "silverfish" would be classified as a concrete, processive, animate noun (41). Other rare singleton and keyword nouns in the passage that also met this classification were then selected at random to create foils. Those selected as foils for "silverfish" using this method were "females," "individuals," and "wasps," as indicated above.

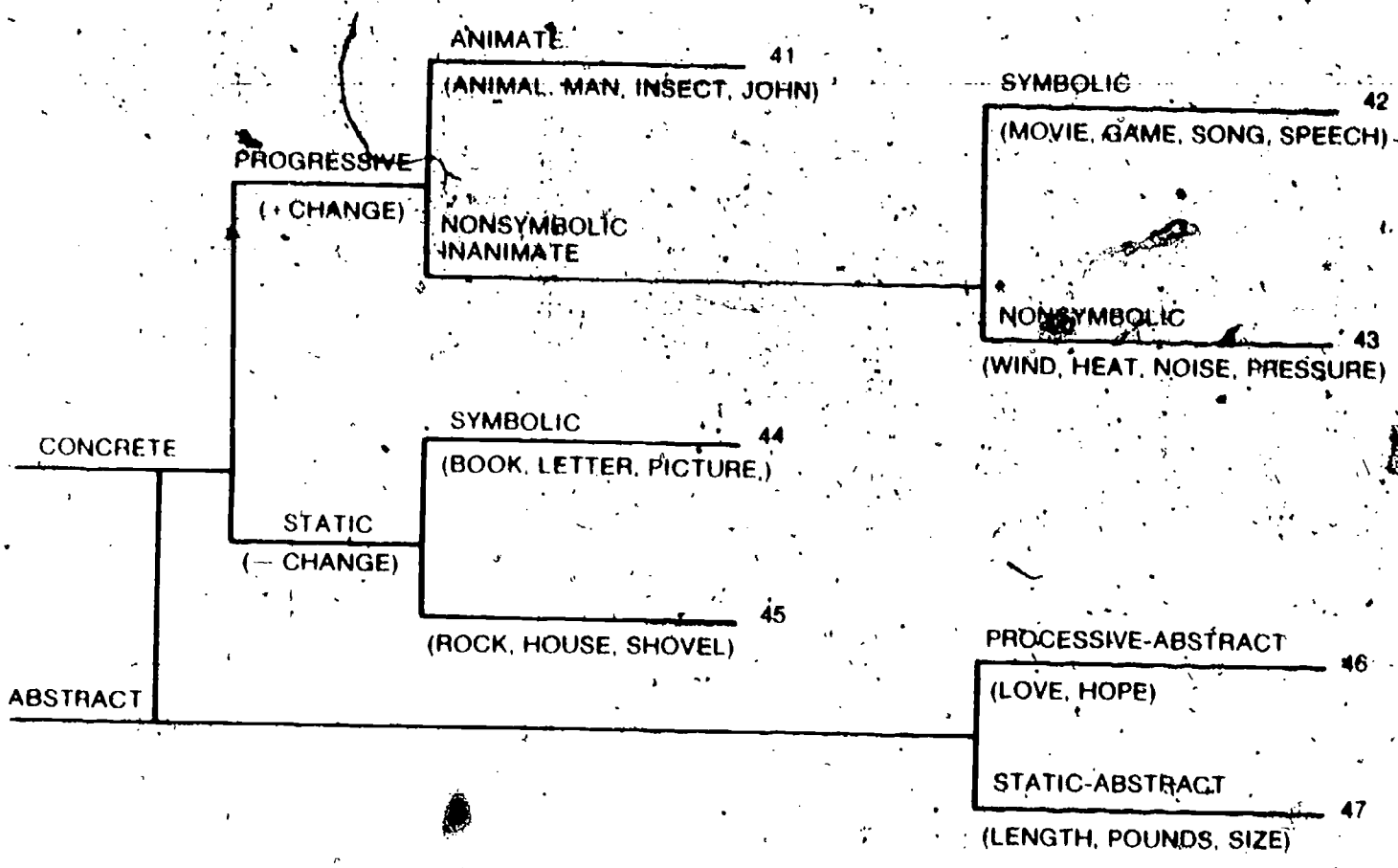

Figure 1. Fredericksen's semantic classification of nouns.

To generate foils for the adjective question words, all rare singleton and keyword adjectives in the prose passage (not just those selected as question words) were classified using semantic differential techniques (Nunnally, 1967, pp. 536-538). In research using these techniques, adjectives are typically classified based on their (1) evaluation (e.g., good or bad), (2) potency (e.g., strong or weak), (3) activity (e.g., fast or slow), and (4) familiarity (e.g., simple or complex). In addition to these four categories, rare singleton and keyword adjectives in the prose passage were classified according to whether or not they could be considered as "technical" words. This latter category is particularly useful in technically-oriented material, particularly for grouping adjectives that relate to a certain noun.

After these adjectives were classified according to these five categories, they were analyzed as to their familiarity, using the Dale-Chall (1948) list of 3000 familiar words. If they were included in that list, they were not used as foils because they were too familiar and, thus, too easy. Approximately 50 adjectives passed this screen and qualified for use as foils. From this group, foils were developed by randomly selecting those having the

same classification as the adjective question words (i.e., as to elevation, potency, etc.). For example, those selected for the rare singleton "pupal" were "nymphal," "parasitic," and "insect" (see appendix).

From the 160 items, eight 20-item test forms were developed. Each test included five items generated from rare singleton nouns; five, from keyword nouns; five, from rare singleton adjectives; and five, from keyword adjectives. In addition, test forms were organized so that each included five items from each of the four item writers, ten items with foils generated informally by the item writers, and ten items with foils generated algorithmically. The internal consistency reliability estimates (Kuder-Richardson Reliability Formula Number 20) averaged .63 for these test forms.

The eight forms were administered to 24 students from the Oregon College of Education before (pretest) and after (posttest) they had studied the prose passage on insect development. For both pretest and posttest, three subjects were randomly assigned to each of the eight test forms; care was taken, however, to ensure that the pretest and posttest forms administered to each student were different.

Average pretest and posttest item difficulties, as determined by the percentages of students who answered the item correctly, were computed for items (1) produced by each of the four writers; (2) derived from each of the four types of question words, and (3) with foils either generated informally by the writers or algorithmically. Also, a nonparametric analysis of variance (ANOVA) (Wilson, 1956) was used to examine differences in item difficulties between (1) the four item writers, (2) the four question word types, (3) the two foil types, and (4) the two test occasions.

Results showed that items based on rare singleton nouns and adjectives and keyword adjectives showed a significant change in item difficulty from pretest to posttest, indicating that such items are useful in learning from the type of prose used in the study. Items derived from keyword nouns, however, produced low quality items, primarily because the sentences they occurred in were usually introductory sentences of a general nature.

The two types of foils proved to be almost equally effective for learning, as evidenced by the similarity in posttest item difficulty. Thus, Roid and Finn concluded that the methods they used for generating foils were feasible. Although foils produced by these methods were somewhat easier than those generated by item writers, they still appeared to produce a significant shift in difficulty from pretest to posttest when instruction was provided between testing sessions.

Finally, the results of the ANOVA showed a strong mean effect for test occasions, which indicates that all types of items were effective for learning. There was also a main effect for word type, which was caused by the easier items derived from keyword nouns, as noted above. Finally, there were two significant three-way interactions: (1) writers by word type by pretest-posttest and (2) writers by foil types by pretest-posttest. The first was caused by variations in item difficulties in items produced by the different writers; and the second, by the fact that one writer generated better foils than the others.

Purpose

The purpose of this study was to extend or replicate the Roid and Finn study. It is expected that the results will form the basis for additional development of algorithmic procedures for generating test questions from prose materials.

APPROACH

Subjects

The eight forms developed in the Roid and Finn study were administered to 249 high school students before (pretest) and after (posttest) they had studied the passage on insect development. For both pretest and posttest, approximately 30 subjects were randomly assigned to each of the eight test forms. Care was taken, however, to ensure that the pretest and posttest forms administered to each subject were different.

Analysis

For purposes of analysis, test results from the earlier study were combined with those obtained in this study. Thus, the total number of subjects was 273 (24 college students and 249 high school students). Since the number of subjects reponding to each test form varied from 27 to 38 on the pretest and from 23 to 33 on the posttest, it was possible to obtain quite stable estimates of item difficulties. A repeated-measures analysis of variance (ANOVA) (4 x 2 x 2 x 2 x 2 factorial design) was used to examine differences in item difficulties between (1) the four item writers, (2) the two parts of speech (adjectives and nouns) of question words, (3) the two types of text frequencies (keyword and rare singletons), (4) the two foil types (writer's choice and algorithmic), and (5) the two test occasions (pretest and posttest).

With 160 items given on two occasions, the analysis had 320 data points, and five replications per cell. The ANOVA, which was conducted on the item difficulties for items in each cell of the design, is useful for determining the "instructional sensitivity" of items. A significant main effect for the pretest-posttest factor would indicate that pretest difficulties were significantly different from posttest difficulties for all items. A significant interaction effect involving the pretest-posttest factor would indicate that certain types of items differed in the pattern of their pretest and posttest difficulties.

RESULTS AND DISCUSSION

ANOVA Results

Table 1, which presents the results of the analysis of variance (ANOVA) of item difficulties, shows that the strongest effect was the main effect for test occasions (R). This finding indicates that, across all types of items, the percentage of subjects getting pretest items correct was lower than the percentage of subjects getting posttest items correct. In other words, most items showed instructional sensitivity. Table 2 shows that pretest item difficulties averaged 47.6 percent across all items; and posttest item difficulties, 74.4 percent. This indicates that the subjects did learn by reading from the passage, even though nearly half were able to guess the correct answer to most questions on the pretest. With four-option multiple-choice items such as those used in this study, excellent items should show pretest difficulties nearer to the level of random guessing (25%).

Two important findings of this experiment were the main effect of part of speech (P) and the interaction of P and the repeated measure (RP), as shown in Table 1. An inspection of Table 3--P and RP interaction effects--reveals that items based on noun question words were significantly easier overall then were items based on adjectives--65.6 vs. 56.3 percent. Also, the difference between pretest and posttest difficulties was greater for nouns than for adjectives (29.5 vs. 24.1%) (untabled), which

Table 1

Repeated Measures Analysis of Variance on Item Difficulties of Items of Each Type

Source	df	F
W (Writers)	3	.25
F (Foil Type)	1	.15
P (Noun vs. Adjective)	1	12.99*
S (Keyword vs. Rare Singleton)	1	2.22
WF	3	.54
WP	3	1.66
FP	1	.27
WS	3	1.29
FS	1	1.33
PS	1	14.21*
WFP	3	.11
WFS	3	.34
WPS	3	.25
FPS	1	.13
WFPS	3	.94
Residual	128	
R (Pretest vs. Posttest)	1	472.03*
RW	3	1.90
RF	1	1.37
RP	1	4.76***
RS	1	2.05
RWF	3	1.74
RWP	3	2.54
RFP	1	3.04
RWS	3	.43
RFS	1	9.25**
RPS	1	20.42*
RWFP	3	1.05
RWFS	3	.61
RWPS	3	2.63
RFPS	1	.11
RWFPS	3	.57
Residual	128	

*p < .001.
**p < .003.
***p < .03.

Table 2

Means and Standard Deviations of Item Difficulties on Pretest and Posttest

Type of Item	Pretest		Posttest	
	Mean	S.D.	Mean	S.D.
Writer (W):				
1	46.8	18.3	71.6	21.4
2	49.5	19.2	72.7	17.4
3	46.7	20.9	75.7	20.2
4	47.2	19.6	77.5	17.2
Foil (F):				
Writer's Choice	46.3	20.1	74.6	19.4
Algorithmic	48.8	18.7	74.2	18.9
Part of Speech (P):				
Noun	50.8	20.1	80.4	14.1
Adjective	44.3	18.2	68.4	21.5
Stem Type (S):				
Keyword	50.4	18.4	75.4	16.4
Rare Singleton	44.8	20.0	73.4	21.5
Test Forms:				
1	47.5	17.3	79.3	20.4
2	48.6	21.4	81.9	17.6
3	44.2	21.2	77.0	19.4
4	51.9	18.9	74.0	21.1
5	51.2	18.3	71.9	16.2
6	46.9	15.3	72.4	13.4
7	37.1	19.8	67.9	21.2
8	53.1	20.5	70.9	21.2
All Items	47.6	19.4	74.4	19.1

Table 3

Means and Standard Deviations of Item Difficulties for Various Interaction Effects

Variable	Repeated Measure (R)					
	Pretest		Posttest		Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
P and RP Interaction Effects						
Noun	50.8	20.1	80.3	14.1	65.6	14.7
Adjective	44.3	18.2	68.4	21.5	56.3	18.1
Average	47.6	19.4	74.4	19.1	61.0	17.1
PS and RPS Interaction Effects						
Noun-based Item:						
Keyword	61.3	15.7	83.5	12.8	72.4	12.3
Rare Singleton	40.4	18.6	77.3	14.8	58.8	13.8
Adjective-based Item:						
Keyword	39.4	14.0	67.4	15.6	53.4	13.4
Rare Singleton	49.1	20.6	69.4	26.3	59.3	21.6
Average	47.6	19.4	74.4	19.1	61.0	17.1
RFS Interaction Effects						
Writer's Choice Foil:						
Keyword	52.5	18.2	75.3	17.5	68.9	16.3
Rare Singleton	40.2	20.2	74.0	21.4	57.1	17.9
Algorithmic Foil:						
Keyword	48.2	18.6	75.6	15.4	61.9	15.8
Rare Singleton	49.3	19.0	72.7	21.9	61.0	18.1
Average	47.6	19.4	74.4	19.1	61.0	17.1

Note. See Table 1 for definitions.

indicates that noun-based items had greater instructional sensitivity than did adjective-based items.

An examination of the PS and RPS interaction effects in Table 3 further reveals the source of the difference between nouns and adjectives in this study. As shown, the average difficulty of items based on keyword nouns is 72.4 percent compared to less than 60 percent for the other types of items. This is because keyword nouns typically occur in introductory sentences that are very general and that address the main topics of the entire passage. For example, in the passage on insect development, the keyword noun "insects" appears in the very first sentence, which happens to be a very general statement--"The life of most insects is short but active." Students can usually answer questions derived from this type of sentence without having to read the prose passage. Also, keyword noun items were relatively easy for subjects to recall on the posttest (average item difficulty of 83.5%), possibly because they were mentioned several times in the passage (see Table 4). This assumption supports Finn's (1977) hypothesis that the information content of rare words is reduced by their high text frequency. Although the fact that keyword adjectives produced the most difficult items (53.4%) appears to be inconsistent with that hypothesis, Table 4 shows that the keyword adjectives occurred fewer times than keyword nouns. Thus, Finn's hypothesis does apply, in that higher text frequency was related to the easiness of items constructed from keywords. With text frequencies of 2 or 3, the keyword adjectives were very close to being rare singletons.

Table 4

Question Words Selected from the Passage and Their Text Frequency

Nouns		Adjectives	
Rare Singleton	Keyword	Rare Singleton	Keyword
Instars	Insect (8)	Plant-feeding	Immature (3)
Cicadas	Insects (20)	Pupal	Incomplete (2)
Silverfish	Metamorphosis (9)	Spine-like	Nymphal (2)
Wasps	Egg (8)	Self-made	Aquatic (2)
Appetites	Adult (8)	Worm-like	Distinctive (2)

Note. The number appearing in parentheses behind keywords represents text frequency.

The rare singleton nouns showed a good pattern of pretest and posttest difficulties. They had the highest average instructional sensitivity--40.4 to 77.3 percent--a difference of 36.9 percent. The rare singleton adjectives were somewhat easier on the pretest and more difficult on the posttest than were the rare singleton nouns.

As shown in Table 1, there was no main effect for writers (W) or foil type (F), nor was there a significant interaction between writers and foil (WF). This result is somewhat surprising in that different writers would be expected to write easier or harder items when they were allowed to choose their own foils.

Table 1 does show one interaction (RFS) involving foil type. The means and standard deviations of item difficulties for that interaction are also included in Table 3. As shown, all of the posttest means are very similar. A Newman-Kuels a posteriori test of the differences between pretest item difficulties in this interaction, however, revealed that, among the items with "writer's choice" foils, the rare-singleton items were more difficult on the pretest than were the keyword items (40.2 vs. 52.5%).

Variance Between Writers

The variability of item difficulties across item writers was examined to determine whether the difficulties of items constructed with "writer's choice" foils varied more across writers than did the difficulties of items constructed with algorithmic foils. It was expected that some writers would choose very difficult foils for a given transformed sentence; and others, easy foils. The algorithmic foils, which were chosen at random from matched groups of similar words from the passage, should be free of any item-writer bias, and, hence, less variable in their effects on item difficulty.

In examining the variability across writers, the focus was on each sentence that was transformed by each writer. As indicated previously, each of the four item-writers produced multiple-choice items (stem and foil) for each of the 20 sentences selected for transformation. It was, therefore, possible to identify four item difficulties for a given combination of sentence and foil technique. For example, for the sentence containing the keyword adjective "immature," the four items generated using the "writer's choice" foil method resulted in pretest difficulties of 38, 65, 52, and 37 percent respectively, and posttest difficulties of 67, 63, 74, and 52 percent. The pretest and posttest variabilities were then calculated across these item difficulties, as shown in Table 5.

After all variances of item difficulties across writers were calculated, they were subjected to a repeated measures ANOVA in which the dependent variables were the natural logarithms of the variances (Scheffé, 1959, p. 83). The design for this analysis was 2 x 2 x 2 x 2 with the following factors: (1) foil type (writer's choice vs. algorithmic), (2) part of speech (noun vs. adjective), (3) stem type (keyword vs. rare singleton question word), and (4) the repeated measure (pretest vs. posttest). Surprisingly, results showed that there were no significant main effects or interactions. For example, even though the average variability of the writer's-choice foil method was 115.31 percent compared to 73.97 percent for the algorithmic foil method, the differences was not statistically significant.

One important limitation of the present study that should be mentioned is that only four item writers were employed. Calculation of variabilities across only four writers is clearly susceptible to the influence of any one of the four item difficulties. With a larger sample of writers, the effects may have been more clearly detectable.

CONCLUSIONS

The concept of using a computer-based algorithm to analyze prose instructional materials and to identify high information words (i.e., those that are rare in American English) appears to be workable. High information nouns or adjectives identified as rare singletons (those occurring only once in a passage) are apparently good candidates for question words. High information adjectives identified as keywords (those occurring more than once in a passage) also appear to be good candidates for question words, providing they occur only two or three times. In contrast, keyword nouns apparently are not good candidates, particularly when they occur in general introductory sentences.

Table 5

Variabilities and Standard Deviations of Item Difficulties

Item Types		Pretest	Posttest	Average
Foil Type:				
Writer's Choice	Var.	131.39	101.21	115.31
	S.D.	11.46	10.06	10.74
Algorithmic	Var.	69.94	78.24	73.97
	S.D.	8.36	8.85	8.60
Part of Speech:				
Noun	Var.	94.45	85.63	89.93
	S.D.	9.72	9.25	9.48
Adjective	Var.	97.30	92.47	94.85
	S.D.	9.86	9.62	9.74
Item Type:				
Keyword	Var.	102.19	87.06	94.32
	S.D.	10.11	9.33	9.71
Rare Singleton	Var.	89.93	90.95	90.44
	S.D.	9.48	9.54	9.51

RECOMMENDATIONS

1. Rare singleton nouns and adjectives and keyword adjectives that occur infrequently in instructional material should be used to select sentences from prose passages for transformation into questions that measure reading comprehension. Keyword nouns should not be used, particularly when they occur in general introductory sentences.

2. Methods of algorithmically generating foils for multiple-choice versions of sentence-derived questions should be further refined and applied in a variety of subject matter areas.

REFERENCES

Dale, E., & Chall, J. S. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11-28.

Finn, P. J. Word frequency information theory and cloze performance: A lexical-marker, transfer-feature theory of processing in reading (Unpublished paper). New York: State University of New York at Buffalo, School of Education, 1977.

Fredericksen, C. H. Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology, 1975, 7, 371-458.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

Roid, G. H., & Finn, P. Algorithms for developing test questions from sentences in instructional materials (NPRDC Tech. Rep. 78-23). San Diego: Navy Personnel Research and Development Center, June 1978. (AD-A056 614).

Scheffé, H. The analysis of variance. New York: Wiley, 1959.

Wilson, K. V. A distribution-free test of analysis of variance hypotheses. Psychological Bulletin, 1956, 53, 96-101.

APPENDIX

THE PROSE PASSAGE USED IN THE EXPERIMENT AND EXAMPLES OF ITEMS PRODUCED FROM TEXT

PROSE PASSAGE USED IN THE EXPERIMENT

4. INSECT DEVELOPMENT

The life of most insects is short but active. Very few insects have a life span of more than a year. By a life span we mean the time from when the egg is laid to when the fully developed adult dies. Let's look at what happens during this period.

All insects develop from eggs. In most cases these eggs hatch outside the body of the female. In the few cases in which the eggs hatch inside the female the young are born "alive." These insects, such as the aphids, are said to be viviparous. (vy vip' ah rus).

Insects that hatch from eggs after they have been laid are said to be oviparous (oh-vip'-ah-rus). Most insects are oviparous. In most cases each egg produces a single immature insect. However, in certain species of parasitic wasps (encyrtids), the egg may produce two or more young.

Most insect eggs are very distinctive. The size, shape, or color of the egg is different, in most cases, for each species of insect. This enables a person who has made a study of these eggs to identify the insect that laid them almost as easily as if he had seen the adult.

Most insect eggs are laid in a place that will provide either protection or food for the young. Protection is especially important to those insects that overwinter in the egg stage. Overwintering means that the adult insect lays its eggs in the late summer or early fall. The eggs then are dormant until the next spring when they hatch. Most of the adults of these species are killed by the first frost. However, the hatching of these eggs in the spring produces new individuals to carry on the species.

Most plant-feeding insects instinctively lay their eggs on plants that the young feed on. This increases the immature insects' chances of survival. If this field of investigation interests you, the study and photography of insect eggs might make a good project.

After reaching the proper stage of development, the egg will hatch. The young insect can use a number of ways to get out of the egg. Some insects

chew their way out. Others have special spinelike structures, called egg-bursters, which cut through the shell. There are some eggs which have special weak spots in them. The young insect escapes from these either by wriggling or by taking in air and bursting the shell with internal pressure.

After the Egg

After hatching, all insects, except the most primitive, go through a series of steps in development. These steps are called ***metamorphosis.*** The word metamorphosis comes from two Greek words; meta, meaning to change, and morpho, meaning form. Therefore, metamorphosis means a change in form. This change in form occurs in two different ways. These two ways are called complete and incomplete metamorphosis. The most primitive insects, such as the silverfish, do not go through metamorphosis. When they hatch they look like their parents in every way except that they are smaller. Their development consists of growing larger and becoming able to reproduce.

Incomplete Metamorphosis

Insects which show this type of metamorphosis have young which look very much like the adults of the species. These immature insects are called nymphs. With the exception of some aquatic species, the principal differences between the nymphs and adults are in size and the presence of wings (see illustration at the right).

Now think back to the description of the phylum to which insects belong, *Arthropoda*. Remember, one of the characteristics of these animals is a hard outer covering called an *exoskeleton*. The exoskeleton is made of a nonliving substance called chitin (ki'-tin). Chitin is hard and stiff and has very little "stretch." Inside the exoskeleton there is very little room for growth.

In order to grow, the nymph must escape this self-made prison. It does this by secreting a new exoskeleton under the old one. When this new skin is complete the old skeleton splits down the

back and the insect walks away and leaves it behind. You have probably seen some of these discarded skins, called casts, on tree trunks.

For a time after the insect discards its old skin, the new exoskeleton is soft. This allows the exoskeleton to expand and make room for further growth.

Each of the periods between molts is called an *instar*. Some nymphs go through as many as eight or more instars before emerging as adults.

Aquatic species that undergo incomplete metamorphosis must go through one more step in development. As nymphs they breathe by means of gills. These gills must be replaced by air-breathing organs in the adult stage. This is done in the last nymphal instar. When it is time for the adult to emerge, the nymph rises to the surface and molts. The fully developed adult steps out of the final nymphal skin with fully developed organs for breathing air.

Complete Metamorphosis

This is the type of metamorphosis that most people are familiar with. Butterflies and moths have complete metamorphosis. There are four distinct stages: egg, larva, pupa, and adult. Since the adult's main activity is producing eggs, and I'm sure you know what these are, we will spend our time studying the larva and pupa.

The larvae's main job in life is to eat and grow. They have huge appetites. Larvae are very different from the adults. They do not have compound eyes, wings, and usually have chewing mouth parts even in those orders where the adults have sucking mouth parts.

A larva may continue to eat and grow all summer. As cold weather approaches, it may build a cocoon and pass into the pupal stage.

Most of these insects pass the winter inside the cocoon. Because no activity is visible at this time, the pupa has been falsely called a "resting stage." Actually a great deal of activity is going on. The wormlike larva is changing into a fully developed adult. When the weather is warm again, this adult emerges from the cocoon, mates, lays eggs, and starts the whole process over again.

Let's Get Together

Most insects reproduce sexually. This means that to have eggs that will hatch, a male and a female of the species must mate. The question is: How do they find each other?

It has been known for years that some of the sounds made by crickets and cicadas were a type of mating call. It is easy to see how these insects get together. But what about the insects that do not make noise; butterflies, for instance?

It has been discovered that the females of these species give off a distinctive odor. This odor is detectable by male insects over great distances. The male follows this scent trail back to the female.

This brings to mind an interesting experiment you might try. A friend of mine once caught a recently emerged female Promethea moth. He put the female in a screen cage and set it outside his window. In less than two hours there were more than twenty males hanging on the outside of the cage. Why don't you try this with other kinds of insects? It would make a great science project.

Science has used the discovery of these odors to help eliminate undesirable insects. It was found that female cockroaches gave off an attractive (to male cockroaches) odor. Scientists have been able to reproduce this scent and have used it to attract males to traps.

Exercises

How Well Did You Read?

1. Name and describe the three types of development insects can go through.
2. What advantage is there in insect eggs being laid on certain plants?
3. What is metamorphosis? What are the differences between complete and incomplete metamorphosis?
4. What processes take place during the growth of insects?
5. Can you think of any advantages to some insects in being born "alive"?

Read A Little More

1. Lemmon, R. S., *All About Moths and Butterflies*, New York: Random House, 1956.

EXAMPLES OF ITEMS PRODUCED FROM TEXT

I. Keyword Noun--Metamorphosis.

a. Text Sentence(s): After hatching, all insects, except the most primitive, go through a series of steps in development. These steps are called metamorphosis.

b. Items (Stem and Foils) Produced by Item Writers:

(1) What are the series of steps in insect development called?

(a) Maturation
(b) Metamorphosis
(c) Symbiosis
(d) Meitosis

(2) What are the steps insects go through in development called?

(a) Metamorphosis
(b) Arthropoda
(c) Larva
(d) Pupa

(3) What are a series of steps in development called?

(a) Reproduction
(b) Larvae
(c) Metamorphosis
(d) Changes

(4) What are the series of steps in insect development called?

(a) Encrytid
(b) Instar
(c) Arthorpoda
(d) Metamorphosis

c. Foils Produced Algorithmically:

Growths
Metamorphosis
Types
Activities

2. Rare Singleton Noun--Silverfish.

a. Text Sentence: The most primitive insects, such as the silverfish, do not go through metamorphosis.

b. Items (Stem and Foils) Produced by Item Writers:

(1) What does not go through metamorphosis? The

(a) Moth
(b) Silverfish
(c) Nymphs
(d) Butterfly

(2) What do not go through metamorphosis? The most primitive insects, such as

(a) Silverfish
(b) Termites
(c) Spiders
(d) Moths

(3) What insects do not go through metamorphosis? The primitive, such as

(a) Eggs
(b) Silverfish
(c) Chitin
(d) Butterflies

(4) The most primitive insects, such as what, do not go through metamorphosis?

(a) Butterflies
(b) Silverfish
(c) Canines
(d) Cicadas

c. Foils Produced Algorithmically:

Silverfish
Females
Individuals
Wasps

3. Keyword Adjective--Immature.

a. Text Sentence: In most cases, each egg produces a single immature insect.

b. Items (Stem and Foils) Produced by Item Writers:

(1) What does each egg produce in most cases? A single

(a) Immature insect
(b) Adult insect
(c) Adolescent insect
(d) Mature insect

(2) What does each egg produce in most cases? A single

(a) Oviparous insect
(b) Nymphal insect
(c) Mature insect
(d) Immature insect

(3) In most cases, what does each egg produce? A single

(a) Dormant insect
(b) Adult insect
(c) Adult insect
(d) Immature insect

(4) What does each egg produce? A single

(a) Immature insect
(b) Mature ubsect
(c) Round insect
(d) Adult insect

c. Foils Produced Algorithmically:

Complete insect
Distinct insect
Immature insect
Incomplete insect

4. Rare Singleton Adjective--Pupal.

a. Text Sentence(s): A larva may continue to eat and grow all summer. As cold weather approaches, it may build a cocoon and pass into the pupal stage.

b. Items (Stem and Foils) Produced by Item Writers:

(1) What may a larva do as the cold weather approaches? Build a cocoon and pass into the

(a) Nymphal stage
(b) Parasitic stage
(c) Pupal stage
(d) Molt stage

(2) As cold weather approaches, a larva may build a cocoon and pass into what?

(a) Infant stage
(b) Adult stage
(c) Butterfly stage
(d) Pupal stage

(3) Into what stage may the larva pass as cold weather approaches and it builds a cocoon? The

(a) Larval stage
(c) Pupal stage
(c) Skeletal stage
(d) Nymphal stage

(4) As cold weather approaches, what may a larva do? Build a cocoon and pass into the

(a) Pupal stage
(b) Hibernation stage
(c) Dormant stage
(d) Resting stage

c. Foils Produced Algorithmically:

Pupal stage
Nymphal stage
Parasitic stage
Insect stage

DISTRIBUTION LIST

Chief of Naval Operations (OP-102) (2), (OP-11), (OP-987H)
Chief of Naval Material (NMAT 08D2)
Chief of Naval Research (Code 450) (3), (Code 452), (Code 458) (2)
Chief of Information (OI-2252)
Director of Navy Laboratories
Chief of Naval Education and Training (00A), (N-5), (N-9)
Chief of Naval Technical Training (Code 016), (Code N-824)
Commander Training Command, U.S. Atlantic Fleet (Code N3A)
Commander, Naval Military Personnel Command (NMPC-013C)
Commanding Officer, Fleet Combat Training Center, Pacific (Code 00E)
Commanding Officer, Naval Education and Training Program Development Center (Technical Library) (2)
Commanding Officer, Naval Education and Training Support Center, Pacific (Code N1B)
Commanding Officer, Naval Health Sciences Education and Training Command (Code 2) (2)
Commanding Officer, Naval Training Equipment Center (Technical Library)
Officer in Charge, Naval Instructional Program Development Detachment, Great Lakes
Officer in Charge, Naval Education and Training Information Systems Activity, Memphis Detachment
Officer in Charge, Central Test Site for Personnel and Training Evaluation Program
Director, Training Analysis and Evaluation Group (TAEG)
Provost, Naval Postgraduate School
Master Chief Petty Officer of the Force, Naval Education and Training Command (Code 003)
Personnel Research Division, Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base
Occupational and Manpower Research Division, Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base
Technical Library, Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base
Flying Training Division, Air Force Human Resources Laboratory, Williams Air Force Base
CNET Liaison Office, Air Force Human Resources Laboratory, Williams Air Force Base
Technical Training Division, Air Force Human Resources Laboratory, Lowry Air Force Base
Advanced Systems Division, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base
Chief, Formal Training Division, Headquarters 34 Tactical Airlift Training Group (MAC), Little Rock Air Force Base
Program Manager, Life Sciences Directorate, Air Force Office of Scientific Research (AFSC)
Army Research Institute for the Behavioral and Social Sciences (Reference Service)
Army Research Institute for the Behavioral and Social Sciences Field Unit--USAREUR (Library)
U.S. Army TRADOC Systems Analysis Activity, White Sands Missile Range (ATTA-SL, Library)
Director, Defense Activity for Non-Traditional Education Support
Secretary Treasurer, U.S. Naval Institute
Science and Technology Division, Library of Congress
Commandant, Coast Guard Headquarters (G-P-1/62)
Commanding Officer, U.S. Coast Guard Training Center, Alameda
Commanding Officer, U.S. Coast Guard Institute
Defense Technical Information Center (12)