ED 163 024 TM 007 732

AUTHOR	Roid, Gale; Finn, Patrick
TITLE	Algorithms for Developing Test Questions from Sentences in Instructional Materials. Interim Report, January-September 1977.
INSTITUTION	Navy Personnel Research and Development Center, San Diego, Calif.
SPONS AGENCY	Advanced Research Projects Agency (DOD), Washington, D.C.
REPORT NO	NPRDC-TR-78-23
PUB DATE	Jun 78
CONTRACT	MDA-903-77-C-0189
NOTE	30p.
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Algorithms; *Complexity Level; Criterion Referenced Tests; Form Classes (Languages); Higher Education; Item Analysis; *Multiple Choice Tests; *Test Construction; *Test Items; Test Reliability; Test Validity
IDENTIFIERS	Item Discrimination (Tests)

ABSTRACT

The feasibility of generating multiple-choice test questions by transforming sentences from prose instructional materials was examined. 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. Items were organized into tests and administered to college students before and after they had studied instructional materials. Results indicated that this item-writing technique was feasible, and that algorithmic methods of generating foils produce items of reasonably good quality. (Author/CTM)

NPRDC TR 78-23 June 1978

ALGORITHMS FOR DEVELOPING TEST QUESTIONS FROM SENTENCES IN INSTRUCTIONAL MATERIALS

Gale Roid
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.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency.

Reviewed by
John D. Ford, Jr.

Approved by
James J. Regan
Technical Director

Navy Personnel Research and Development Center
San Diego, California 92152

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
NPRDC TR 78-23

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
ALGORITHMS FOR DEVELOPING TEST QUESTIONS FROM SENTENCES IN INSTRUCTIONAL MATERIALS

5. TYPE OF REPORT & PERIOD COVERED
Interim Report
January-September 1977

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Gale H. Roid
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
627094N-RPA.3354

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

12. REPORT DATE
June 1978

13. NUMBER OF PAGES
29

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 feasibility of generating multiple-choice test questions by transforming sentences from prose instructional materials was examined. 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.

Items were organizec into tests and administered to subjects before and after they had studied instructional materials. Results indicated that this item-writing technique was 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.

This interim report describes the beginning phasea of a contractual effort aimed at examining the qualities of test questions written from a variety of methods. Subsequent reports will deal with further comparisons of various item writing methodologies and the development of a handbook on item writing technologies associated with criterion-referenced testing.

Appreciation is expressed to Dr. Tom Haladyna of the Teaching Research Division, Oregon State System of Higher Education, a research associate in this effort, and to Dr. John R. Bormuth of the University of Chicago, a consultant for the project.

The Contracting Officer's Technical Representative was Dr. Pat-Anthony Federico of this Center.

J. J. CLARKIN
Commanding Officer

C

CONTENTS

Page

INTRODUCTION 1

Problem 1
Background 1
Objective 4

APPROACH 5

Item Development 5
Algorithmic Foil Generation 6
Test Construction and Administration 7
Analyses 8

RESULTS 9

Average Item Difficulty and Instructional Sensitivity 9
Analysis of Average Item Difficulty 11

CONCLUSIONS 13

RECOMMENDATIONS 15

REFERENCES 17

REFERENCE NOTES 29

APPENDIX--THE PROSE PASSAGE USED IN THE EXPERIMENT AND EXAMPLES OF ITEMS PRODUCED FROM TEXT A-0

DISTRIBUTION LIST

LIST OF TABLES

1. Question Words Selected 5

2. Average Item Difficulty and Instructional Sensitivity 10

3. Results of a Nonparametric Analysis of Variance on Item Difficulties for Items in Each Category 11

INTRODUCTION

Problem

Measurement theorists have argued convincingly that the current crisis in education stems from the lack of a scientific basis for writing achievement test questions, or items. This crisis has been intensified by an increased public demand for accountability in education and by interest in the use of tests for selection, placement, advancement, certification, and other important decisions that deeply affect people's lives. Although it is reasonable to expect that such decisions would involve reliable and appropriate tests, test specialists currently must work without the aid of a systematic technique for writing test items. Instead, for both criterion-referenced tests (in which an individual's performance is compared to a standard rather than to that of other individuals) and for traditional norm-referenced tests, they must rely on their intuitive skills or on those of experts to assess questions' merits.

Even when item writers are given learning objectives that describe what is to be learned in terms of expected student performance under specified conditions and standards, they will not necessarily generate the same items or even items of similar quality. Current military guidelines for designing criterion-referenced tests for use in instructional systems (Swezey & Pearlstein, 1974) refer to the "writing of test items for each learning objective," but do not provide detailed suggestions for writing such items. Item-writing methods are needed that are (1) based on a logically and precisely defined relationship between the text and the test items written to assess learning from that text, (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.

Use of 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

Although theories and suggestions have been published concerning new item-writing methods, little specific research has been conducted to determine either the technical quality of items written by such methods or the feasibility of their widespread use in education and training. Only a handful of civilian research studies, most of which are currently unpublished, have examined the technical and measurement qualities of the new item-writing methods, such as those capable of being produced algorithmically. If these methods are to be used in military training and to reshape the everyday practices of educational testing in the United States, they must have a strong research base.

There is an even more practical reason for interest in algorithmic methods of writing test questions: When students are to be retested aeveral times, particularly when using instructional systems that involve the mastery learning model (Bloom, 1968), multiple test forms must be provided that are equivalent in both content coverage and difficulty. Although such test forms could be assessed and revised through field tests, much time and energy could be saved if forms of near equivalency could be produced algorithmically.

Roid and Haladyna (1978), in comparing item-writing techniques (e.g., Millman, 1974; Bormuth, 1970), found that one of two item writers produced consistently more difficult test items from the same learning objectives. The resulting differences in test difficulty would have serious implications for the criterion-referenced uses of such tests (e.g., those affecting pass-fail decisions).

Anderson (1972, pp. 151-159) proposed various item-writing methods to test the learning of concepts and principles. These methods rely on an analysis of examples and nonexamples of a concept or a principle and usually go beyond the verbatim wording used in the instructional materials. Tiemann, Kroeker, and Markle (Note 1) have devised plans for sampling examples and nonexamples of concepts in both teaching and testing settings.

Bormuth (1970) proposed operationally defined item-writing rules for transforming segments of prose material to obtain items that test recall of such material. Specifically, he proposed rules for deriving items from sentences, and from the relationships between sentences (pp. 39-55). An example of sentence-derived items are those produced by the "wh-transformation," which requires the writer to inspect all sentences in the instruction and to substitute a "wh-pro" word such as who, what, or where for, say, the subject of each sentence. For instance, "The boy rode the horse" could be transformed to "Who rode the horse?" Items derived by this method are particularly useful because they can be written to cover each part of a sentence and tailored to either the multiple-choice or fill-in format. Sentence-derived items can also result through the use of paraphrasing; that is, by replacing substantive words in a sentence with others having the same meaning.

Items can be derived from the relationships between sentences by questioning the cause of a described action or result. For instance, the sentences "Jim hurt his foot," "He was cleaning his gun," and "His gun accidently fired" can be examined for implied causation, resulting in the question "What caused Jim's hurt foot?"

Finn (1975) extended Bormuth's work by developing a question-writing algorithm for learning from prose. The principle steps in this algorithm are described in the following paragraphs.

1. Computer Analysis of Passage or Test. The passage or text is analyzed by keypunching all words and entering them in a computer program that (a) counts the number of times that each word appears in the passage (text frequency) and (b) calculates its standard frequency index (SFI), which is a numerical estimate of how often the word appears in a large corpus (five million words) of American English (Carroll, Davies, & Richman, 1971). The SFI ranges from 88.6 for the word "the" to 02.5 for the word "incarnation" (i.e., the average student is likely to encounter the word "the" once in every 10 words of his schoolbook reading and the word "incarnation" less often than once in every billion words.

2. Identification of Candidate Sentences for Transformation into Items. Words having a low SFI--that is, they are relatively rare in American English--are called high information words. The sentences in which these words appear

can be regarded as candidates for transformation into questions that tap important information in the passage.

3. Selection of High Information Words for Use as Question Words. High information words usually are difficult for subjects to guess if they are deleted from a prose passage, which is the method used in cloze tests (Culhane, 1970). In such tests, segments of prose are presented to a subject, usually with every fifth word deleted, and he is tasked to supply the missing words. The ease with which he supplies a missing word is a measure of the amount of information it provides.

Finn (Note 2) found that the cloze easiness of a word can be predicted by the two indices derived from computer analysis of a passage; that is, word frequency and SFI. A word having a low SFI is typically high in information. However, if this word appears frequently in the passage, its information value will be diminished because subjects will supply it more easily in a cloze test following reading of the passage. In other words, repetition of words, even if they are rare in American English, lowers their information value. Therefore, Finn concluded that good candidate question words must have a low SFI and must occur only once in a prose passage.

Not all parts of speech--even if they meet the above criteria--are equally good candidates for question words. Verbs and adverbs pose particular problems. For example, the sentence, "Finn echoed the concern of Bormuth," when transformed to "What did Finn do to the concern of Bormuth?" is clumsy and less important than "Who echoed the concern of Bormuth?" After considerable effort to produce questions from verbs and adverbs, the authors of this report concluded that the most promising question words are adjectives, nouns, or phrases including an adjective or a noun.

Adjectives and nouns can be further classified by type. For example, either may be part of a noun phrase, and nouns may be possessive. If an algorithm is to be fully defined, then, the classifications of the question words within parts of speech must be specified to eliminate ambiguity for the item writer who selects the words.

4. Sentence Analysis. Once a question word has been selected, the sentence in which it occurs is analyzed or diagrammed to identify its important parts (e.g., subject, verb, and object). This procedure is advantageous for two reasons. First, parts of speech that are least promising for question words (i.e., explicatives, functional verbs, articles, and prepositions) either appear as parts of phrases or not at all. Second, the number of questions possible for a given sentence becomes a function of the number of case phrases and nonzero verbs in the sentence rather than the number of words.

5. Sentence Transformation. The next step is to transform the sentence into a question by replacing the question word, usually an adjective, a noun, or a phrase including an adjective or a noun, with a wh-word. Where several wordings are possible, an attempt is made to stay as close as possible to the wording of the original sentence. Sentences may also be transformed by replacing pronouns with their appropriate nouns and references

to previous sentences with clauses or phrases from those sentences However, this method does not produce 100 percent agreement among item writers.

6. Algorithmic Generation of Foils (response alternatives). The first step in an algorithmic generation of foils is to classify the correct alternative so that possible foils can be obtained from a list of words similarly classified. The most logical source of foils would seem to be the prose passage itself but, in some cases, published lists of words (e.g., Carroll et al., 1971) may be useful.

Objective

The objective of the present effort was to refine procedures for choosing question words for use in wh-transformations of instructional sentences and for algorithmically generating multiple-choice foils. Multiple-choice testing is the most common testing method used in education and training.

APPROACH

Item Development

A prose passage on insect development, which was written for approximately the high school level, was selected for use in this study. This passage is provided in the appendix. Items (stem and foils) to test learning from this passage were then developed using the following procedure:

1. All of the words in the passage were keypunched into a computer program to determine their standard frequency index (SFI) and text frequency. Nouns and adjectives having an SFI of 60 or less were identified, since they appeared to be the best candidates for question words. These nouns and adjectives were then further classified to identify those that (1) appeared only once in the text and, (2) had a high text frequency. For the remainder of this report, these two classifications are referred to as rare singletons and keywords.

2. Twenty sentences were selected for transformation into items. Five of these sentences included rare singleton nouns; five, keyword nouns; five, rare singleton adjectives; and five, keyword adjectives. These nouns and adjectives are listed in Table 1.

Table 1

Question Words Selected

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.

3. The selected sentences were transformed (using the wh- method) into multiple-choice items by four item writers (Author Finn and three graduate students from the State University of New York at Buffalo). After working as a team to ensure that items produced were similar, the writers produced items independently. For each of the 20 sentences selected, each writer produced two items: The stems for the two items were identical but the foils or alternatives for one item were generated informally by the writer

and those for the second item, by an algorithmic method. 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 stem: "The most primitive insects, such as what, do not go through metamorphosis?" The first item formed using this stem included foils produced informally by the author, in this case:

1. Butterflies
2. Silverfish
3. Canines
4. Cicadas

The second item included foils generated algorithmically, in this case:

1. Silverfish
2. Females
3. Individuals
4. Wasps

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

Algorithmic Foil Generation

In generating foils algorithmically, the writers experimented with a method based on the Word Frequency Index (Carroll et al., 1971), which provides the SFIs for more than five million words. Question words (e.g., silverfish) were located in the index and those in the index having similar SFIs were located for possible use as foils. However, the index proved to be an unacceptable source for this particular application; thus, an algorithmic method of foil construction was developed that extracted foils from the prose passage itself, and variations of that algorithm were developed for nouns and for adjectives.

The rare singleton and keyword nouns selected as question words were classified semantically using the method developed by Fredericksen (1975), which is shown in Figure 1. For example, 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.

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.

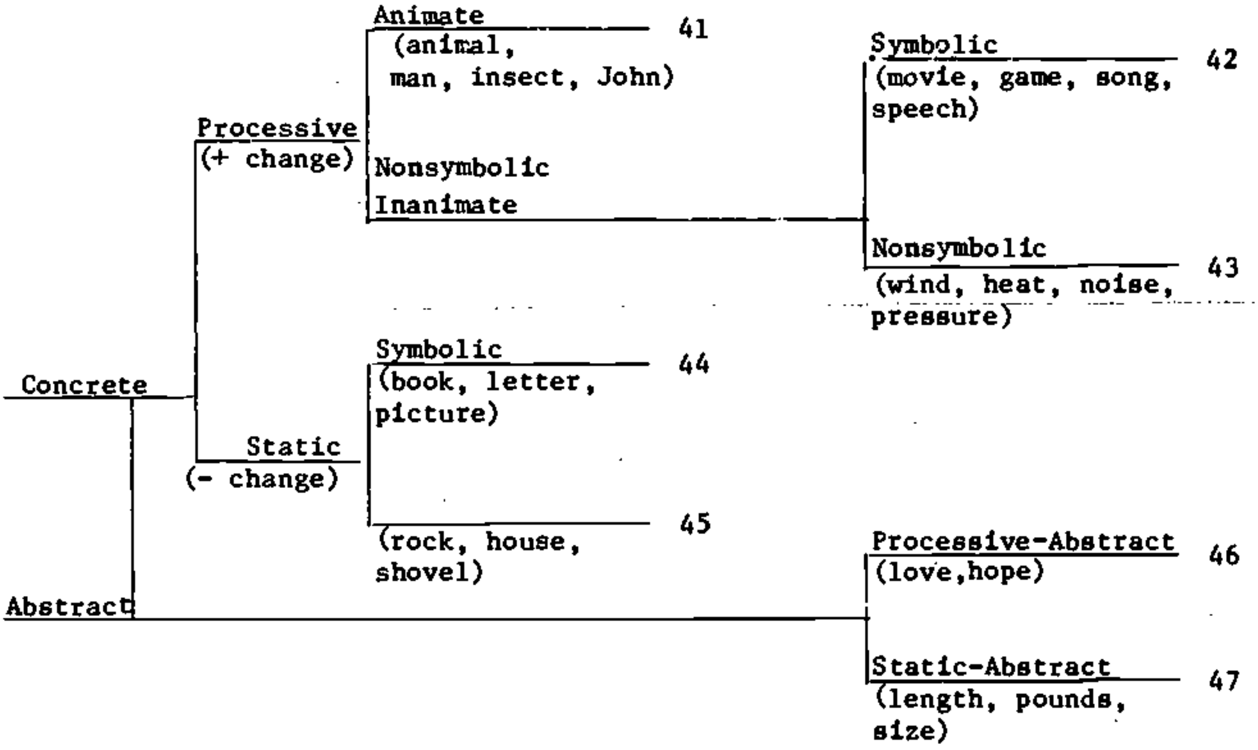


Figure 1. Fredericksen's semantic classification of nouns.

After these adjectives were classified according to the five categories noted above, they were subjected to an analysis of familiarity, using the Dale-Chall (1948) list of 3000 familiar words. If they were included in that list, they were not considered for use as foils because they were too familiar and, thus, too easy. Approximately 50 adjectives passed this screen and qualified for use as foils. Foils for adjective question worda were then developed by randomly selecting those having the same classification (i.e., as to elevation, potency, etc.). For example, those selected for the rare singleton "pupal" were "nymphal," "parasitic," and "insect" (see appendix).

Test Construction and Administration

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, 10 items with foils generated informally by the item writers, and 10 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; however, care was taken to ensure that the pretest and posttest forms administered to each student were different.

Analyses

Average pretest and posttest item difficulties, as determined by the percentages of students who answered the item correctly, were computed for items in the following categories: (1) those produced by each of the four writers, (2) those derived from each of the four types of question words, and (3) those with foils either generated informally by the writers or algorithmically. It was hypothesized that items generated from rare singleton nouns and adjectives would provide the best instructional sensitivity, as determined by the difference between their pretest and posttest item difficulties.

Due to possible fluctuations in item difficulty because of the small sample size, 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.

With 160 items administered on two occasions, the analysis had 320 data points and five replications per cell. The nonparametric ANOVA is based on identifying the number of item difficulties that fall above or below a grand median; thus, contingency tables were created to display the number of observations falling above or below the median in each cell of the factorial design, as suggested by Wilson (1956). The chi-square statistic for the contingency table, created by using all four factors in the design, was then decomposed into sources of variation in the same manner that a total sum-of-squares is decomposed in a parametric ANOVA. The decomposition of chi-square was shown originally by Rao (1952, pp. 192-205).

The ANOVA is also useful for determining items' instructional sensitivity: 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

Average Item Difficulty and Instructional Sensitivity

Table 2, which provides average item difficulty and instructional sensitivity, indicates that items derived from rare singleton nouns showed a good pattern of pretest and posttest difficulty (56.2 to 88.3%), and had the highest mean instructional sensitivity (32.1%). Items derived from rare singleton adjectives showed a pattern of average item difficulties similar to that of rare singleton nouns (54.4 to 79.3%); however, these items were somewhat more difficult than the former on the posttest. Also, the mean instructional sensitivity for rare singleton nouns was not as high as that for keyword adjectives (24.9 vs. 29.6%). Thus, the hypothesis that rare singleton nouns and adjectives would provide the best instructional sensitivity was only partly supported.

Table 2 also shows that items derived from keyword nouns were significantly easier on the pretest than were items derived from the other question words. An examination of the text sentences in which these words appeared showed that they were typically introductory and, thus, very general. For example, the keyword noun "insects" appears in the very first sentence: "The life of most insects is short but active." Items derived from such general statements usually concern common knowledge that students can answer correctly without having to read the prose passage. Further, items based on keyword nouns were easier on the posttest than the others, although not to a significant degree. This finding supports the hypothesis (Finn, Note 2) that the information content of words (even if they are rare in American English) is reduced by their high text frequency. As shown in Table 1, keyword nouns used in this study had a text frequency ranging from 8 to 20.

Keyword adjectives produced the most difficult items on the posttest, a finding which is not consistent with the above hypothesis. The reason for this apparent inconsistency is shown in Table 1: With text frequencies of two or three, the keyword adjectives were very close to being rare singletons.

The two types of foils proved to be almost equally effective for learning, as evidenced by the similarity in posttest item difficulty. However, those that were informally generated by the item writers were considerably harder on the pretest (i.e., students were not able to guess the correct answer as often when such foils were used), and had a much higher instructional sensitivity than algorithmically generated foils (30.5 vs. 19.4). This is understandable, since any automated method inevitably will produce some implausible foils. A skilled item writer, on the other hand, can choose foils that fit the meaning and semantic qualities of the item stem and the correct foil.

Table 2

Average Item Difficulty and Instructional Sensitivity

	Pretest		Posttest		Instructional Sensitivity
Item Category	Mean	S.D.	Mean	S.D.	(Posttest Minus Pretest)
Item Writers					
#1 (N = 5)	62.9	37.8	81.9	28.5	19.0
#2 (N = 5)	65.0	36.4	85.8	28.3	20.8
#3 (N = 5)	49.5	36.7	82.9	30.9	33.4
#4 (N = 5)	57.7	33.8	83.7	28.1	26.0
All Writers (N = 20)	58.8	36.2	83.6	28.9	24.8
Type of Word					
Rare Singleton Noun (N = 5)	56.2	34.4	88.3	21.5	32.1
Keyword Noun (N = 5)	77.1	31.2	89.6	25.0	12.5
Rare Singleton Adjective (N = 5)	54.4	41.3	79.3	32.2	24.9
Keyword Adjective (N = 5)	47.5	31.9	77.1	33.6	29.6
All Types of Words (N = 20)	58.8	34.7	83.6	28.1	24.8
Type of Foil					
Writer Generated (N = 10)	54.7	37.4	85.2	29.0	30.5
Algorithmically generated (N = 10)	62.9	35.0	82.3	28.5	19.4
Both Types of Foil (N = 20)	58.8	36.2	83.8	28.8	24.9

Analysis of Average Item Difficulty

The results of the nonparametric analysis of variance on average item difficulty are presented in Table 3. The main effect for test occasions (D) was str ngest, which indicates that, across all types of items, a higher percentage of students answered items correctly on the posttest than the pretest (83.5 vs. 58.8% on Table 2). In other words, most items showed instructional sensitivity: the students did learn from reading the passage. Further, the overall pretest item difficulty of 58.8 percent indicates that over half the students were able to guess the correct answer to most questions without reading the passage. Thus, the items developed could not be rated "excellent"; with four-alternative, multiple-choice items, such as those used in this study, "excellent" items should show pretest difficulties nearer to the level of random guessing; that is, 25 percent.

Table 3

Results of a Nonparametric Analysis of Variance on Item Difficulties for Items in Each Category

Source of Variation	Chi-Square	df
A (Writers)	2.51	3
B (Word types)	16.32	3*
C (Foil types)	.31	1
D (Pretest vs. Posttest)	45.53	1*
AB	8.24	9
AC	1.28	3
AD	2.86	3
BC	2.07	3
BD	2.25	3
CD	3.71	1
ABC	7.97	9
ABD	18.29	9**
ACD	8.40	3**
BCD	4.01	3
ABCD	12.45	9
Total	134.20	63

*p < .001
**p < .05

There was also a main effect for word type (B). This effect was caused by the fact that items derived from keyword nouns were significantly easier on the pretest than other items. The reason for this was discussed previously.

As shown, there were no main effects for writers (A) or foil types (C) or significant two-way interactions. However, there were two significant three-way interactions: (1) ABD (writers by word type by pretest-posttest) and (2) ACD (writers by foil types by pretest-posttest). Inspection of the item difficulties in each cell for the ABD interaction indicated the following variations between writers:

1. Writers #2 and #4 wrote keyword noun items that were much easier for students to guess correctly on the pretest than those written by Writers #1 and #3.

2. Writer #2 wrote rare singleton noun items that were much easier for students to answer correctly on the posttest than did the other writers.

3. Writer #4 wrote "excellent" rare singleton adjective items, as indicated by the high instructional sensitivity they showed from pretest to posttest.

Examination of the ACD interaction revealed that Writer #3 generated excellent foils, as evidenced by the high instructional sensitivity items with such foils showed from pretest to posttest. A comparison of foils generated by Writer #3 with those generated by other writers showed that he had selected foils that were more (1) logically related to the passage, (2) difficult, and (3) semantically parallel to the correct answer.

Although the effects of the significant three-way interactions found in this study were not as strong as the main effects for test occasion or word type, they do suggest two important possibilities:

1. The skill of item writers will vary to the extent that a good item writer can produce foils that are better than those produced algorithmically.

2. An algorithmic foil-generating method can smooth out differences between item writers with different capabilities.

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.

The methods used in this study to generate foils algorithmically for multiple-choice versions of sentence-derived items appear to be feasible. Although foils generated in this manner may be somewhat easier than those generated by item writers, they still appear to produce significant instructional sensitivity--a shift in difficulty from pretest to posttest when instruction is provided between testing sessions.

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

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Bloom, B. S. Learning for mastery. Evaluation comment, UCLA, Vol. 1. No. 2., May 1968.

Carroll, J. B., Davies, P., & Richman, B. Word frequency book, Boston: Houghton-Mifflin, 1971.

Cronbach, L. J., & Bormuth, J. R. On the theory of achievement test items. Psychometrika, 1970, 35, 509-511. (Book Review)

Culhane, J. W. CLOZE procedures and comprehension. The Reading Teacher, 1970, 23, 410-413.

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

Finn, P. J. A question writing algorithm. Journal of Reading Behavior, 1975, 4, 341-367.

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

Millman, J. Criterion-referenced measurement. In Popham, W. J. (Ed.) Evaluation in education: Current applications. Berkeley, CA: McCutchan Publishing Company, 1974.

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

Rao, C. R. Advanced statistical methods in biometric research. New York: Wiley, 1952.

Roid, G. H., & Haladyna, T. A comparison of objective-based and modified-Bormuth item writing techniques. Educational and Psychological Measurement, Spring, 1978.

Swezey, R. W., & Pearlstein, R. B. Developing criterion-referenced tests. Reston, VA: Applied Science Associates, 1974.

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

REFERENCE NOTES

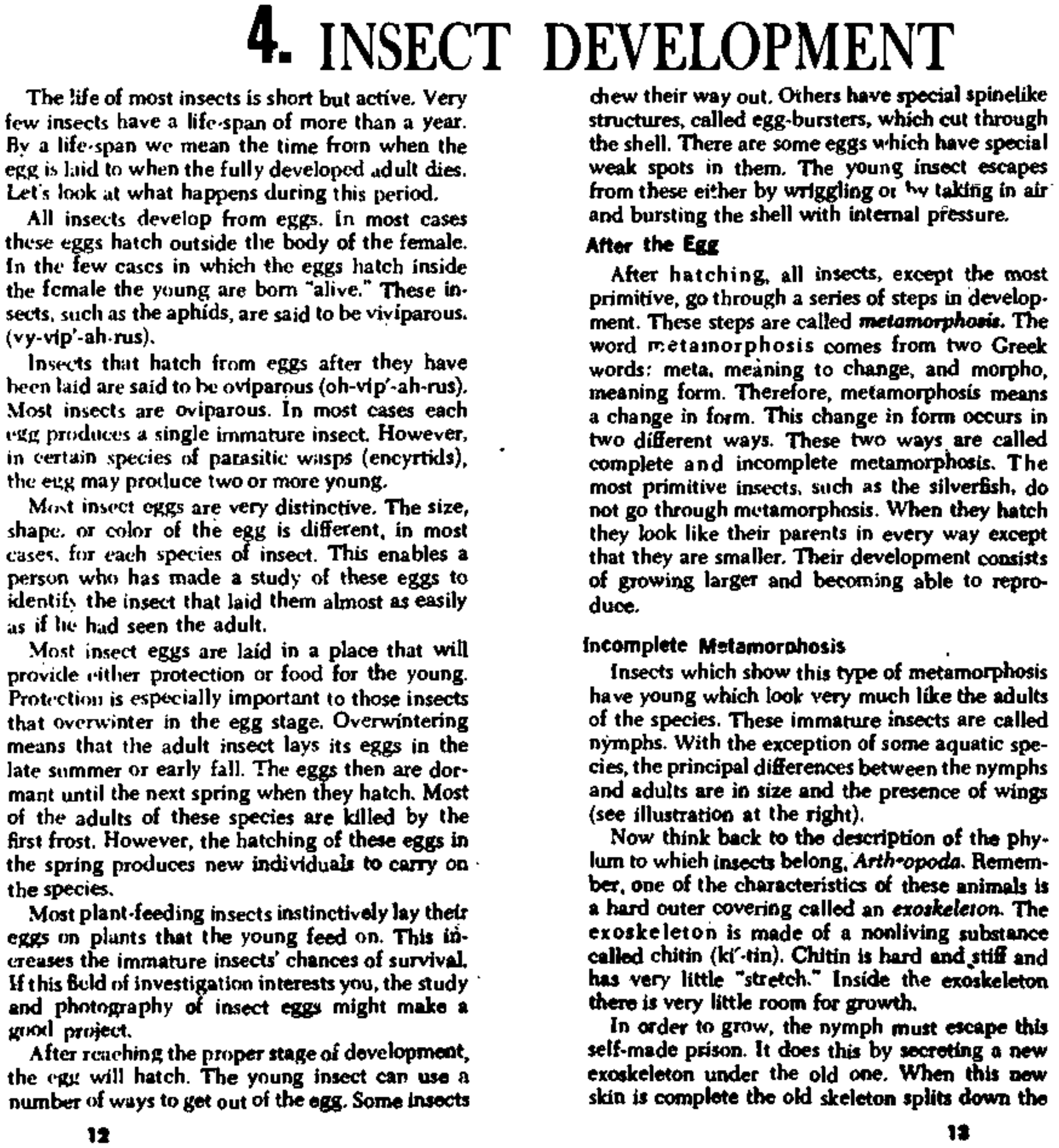
1. Tiemann, P., Kroeker, L. P., & Markle, S. M. *Teaching verbally-mediated coordinate concepts in an on-going college course*. Paper presented at the meetings of the American Educational Research Association, New York, April 1977.

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

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 (kī'-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

Note. Special permission granted by What Insect Is That? published by Xerox Education Publications, (c) 1965 Xerox Corp.

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.

Note. Special permission granted by What Insect Is That? published by Xerox Education Publications, (c) 1965 Xerox Corp.

EXAMPLES OF ITEMS PRODUCED FROM TEXT

1. 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 (c) Symbiosis
         (b) Metamorphosis (d) Meitosis

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

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

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

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

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

         (a) Encrytid (c) Arthorpoda
         (b) Instar (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 (c) Nymphs
         (b) Silverfish (d) Butterfly

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

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

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

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

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

(a) Butterflies (c) Canines
(b) Silverfish (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 (c) Adolescent insect
(b) Adult insect (d) Mature insect

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

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

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

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

(4) What does each egg produce? A single

(a) Immature insect (c) Round insect
(b) Mature ubsect (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 (c) Pupal stage
(b) Parasitic 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-987H), (OP-991B)
Chief of Naval Personnel (Pers-10c), (Pers-2B)
Chief of Naval Material (NMAT 08T244)
Chief of Naval Research (Code 450) (4)
Chief of Information (OI-2252)
Director of Navy Laboratories
Chief of Naval Education and Training (N-5)
Chief of Naval Technical Training (Code 015), (Code 016)
Chief of Naval Education and Training Support
Chief of Naval Education and Training Support (001A), (N-5)
Commanding Officer, Naval Training Equipment Center (Technical Library)
Director, Training Analysis and Evaluation Group (TAEG)
Director, Defense Activity for Non-Traditional Education Support
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
Technical Training Division, Air Force Human Resources Laboratory, Lowry Air Force Base
Flying Training Division, Air Force Human Resources Laboratory Williams Air Force Base
Advanced Systems Division, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base
Program Manager, Life Sciences Directorate, Air Force Office of Scientific Research (AFSC)
Army Research Institute for the Behavioral and Social Sciences
Science and Technology Division, Library of Congress
Coast Guard Headquarters (G-P-1/62)
Secretary Treasurer, U. S. Naval Institute
Defense Documentation Center (12)