DOCUMENT RESUME

ED 133 332 TM 005 562

AUTHOR Roid, G. H.; Haladyna, Thomas M.
TITLE A Comparison of Objective-Based and Modified-Bormuth Item Writing Techniques.
PUB DATE [Apr 76]
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; *Comparative Analysis; Complexity Level; *Methods; *Test Construction
IDENTIFIERS *Test Item Writing

ABSTRACT

Two techniques for writing achievement test items to accompany instructional materials were contrasted: (1) generating items from statements of instructional objectives, and (2) generating items from rules for transforming instructional statements (adapted from Bormuth). Items of each type were written by two experienced item writers. Subjects were given tests employing these items before and after reading a programmed booklet. One item writer was found to produce consistently more difficult test items regardless of the technique used. This result supports the contention that objective-based item writing results in items of varying quality, but is in conflict with the hypothesis that the rule-generation technique eliminates "subjectivity" in item writing. The need for further investigation of fully-automated, linguistic-based rules for item writing is suggested. (Author)

ABSTRACT

A Comparison of Objective-Based and Modified-Bormuth Item Writing Techniques[1]

G. ". Roid and Thomas M. Haladyna
Teaching Research Division
Oregon State System of Higher Education

Two techniques for writing achievement test items to accompany instructional materials were contrasted, (a) generating items from statements of instructional objectives, and (b) generating items from rules for transforming instructional statements (adapted from Bormuth). Items of each type were written by two experienced item writers. Subjects were given tests employing these items before and after reading a programmed booklet. One item writer was found to produce consistently more difficult test items regardless of the technique used. This result supports the contention that objective-based item writing results in items of varying quality, but is in conflict with the hypothesis that the rule-generation technique eliminates "subjectivity" in item writing. The need for further investigation of fully-automated, linguistic-based rules for item writing is suggested.

[1]Paper presented at the meeting of the American Educational Research Association, San Francisco, April, 1976.

A Comparison of Objective-Based and Bormuth Item-Writing Techniques

G. H. Roid and Thomas M. Haladyna
Teaching Research Division
Oregon State System of Higher Education

Bormuth (1970) has suggested that achievement test items should be constructed by using operationally defined item writing techniques, so that a precise description of what has been learned and measured can be made. These techniques involve the algorithmic transformation of sentences from prose instruction into test items and allow for automated item generation. Hively (1974) has proposed the concept of item forms for generating domain-referenced test items, a concept similar to Bormuth's item writing rules. Anderson (1972), Millman (1974), and others have reiterated the need for item writing rules and Anderson has emphasized the importance of insuring that test items contain wording or examples different from those used in instruction in order to truly test comprehension.

Bormuth (1970) has contrasted traditional methods of item construction (as represented by the methods of Bloom, Thorndike, and Hagan, etc.) with item construction using operational definitions or rules. Operational item writing rules for achievement test items are a series of directions which tell an item writer how to rearrange segments of the instruction to obtain the items of that type. A simple example would be "subject deletion" items which would be written using a detailed rule summarized as follows: "Inspect all sentences in the instruction, replacing a "wh-pro" word such as who, what, or where for the appropriate subject each sentence. For instance, "The boy rode the horse" would be transformed to: "Who rode the horse?".

In contrast to operationally defined techniques, Bormuth (1970, p.10) and Millman (1974, p. 325) have suggested that the use of instructional objectives or other traditional item writing methods allow the item writer so many options that no two item writers could be expected to produce comparable tests.

However, as Cronbach (1970) commented, complete and useful guidelines for the average test constructor were not provided in the work of Bormuth (1970). In fact, unless one is a competent linguist versed in transformational grammar, the implementation of Bormuth's method for tests of prose learning will be difficult indeed. The complexity of the transformations required to make sentences into test items is reflected in the 82-step algorithm proposed by Finn (Note 1), a former student of Bormuth's. Also, Hambleton, et al., (Note 2, p. 15) called for a compromise method intermediary to automated item generation and the use of objectives, because "it does not appear likely that the notion of specifying content via the use of item generation rules will be applicable to many subject areas." For these reasons and because of the obvious importance of Bormuth's suggestions, the present study was designed to examine the effectiveness of using rules for item writing which are similar to, but not as rigorously automated as those suggested by Bormuth. The item-writing rules of the present study were used to transform sentences taken directly from an instructional booklet, and used keyword deletion and wh-pronouns in syntactic transformations as suggested. However, the rules were written so that most item writers unfamiliar with linguistics could still implement them.

In addition, to a comparision of objective-based and rule-generated item methods, the present study examined the contention of Anderson (1972) that

paraphrasing of words from the instructional text for use in test items more adequately assesses comprehension than using verbatim words from the text. Specifically, the present study tested the hypothesis that items with verbatim wording would be easier than items with paraphrased wording.

In summary, the study contrasted the following:

1. Items written from statements of the instructional objectives of a programmed instructional booklet vs. items written from item writing rules for the same booklet.
2. Items designed to measure verbatim recall vs. comprehension (comprehension measured by items which use paraphrasing of instructional statements or examples not given in the instruction).
3. Items written using each of the above strategies by <u>two different</u> experienced item writers.

The major question under investigation was "Do different item writers produce test items which are more similar (statistically) if they use operational item writing rules than if they use instructional objectives as a guide."

Method

<u>Subjects.</u> Seventy-two dental students from the University of Oregon Health Science Center, School of Dentistry served as subjects. They were all enrolled in a second-year course in Crown and Bridge Techniques. The experiment was conducted as part of the regular course work and an instructional booklet used in the experiment was required reading.

<u>Instruments.</u> Four parallel test forms of 48 items each were constructed to measure learning from the booklet. Items of each type were written by each of two item-writers:[2]

---

[2]Item writers were the authors, G. H. Roid and Thomas M. Haladyna.

(a) objective-based items

(b) items generated from rules for transforming instructional statements

(c) verbatim recall (recognition of multiple choice alternative), and

(d) comprehension items written by paraphrasing instructional statements or using new example.

Item writing rules were of two types, keyword deletion and syntactical transformation. The rules for verbatim items were written as follows:

Rule 1. Verbatim-Deletion. A keyword or phrase is deleted and replaced by a blank. The word or phrase is included as one of four alternatives in a multiple-choice format. Except for deletion, the exact words from the instructional statement being used are retained.

Rule 2. Verbatim-Syntactical Transformation. Also might be called a "wh-transformation". A keyword or phrase is deleted from an instructional statement and replaced by a phrase such as "which one of the following", or "what is...". The keyword or phrase is included as one of four alternatives in a multiple-choice format. Except for the addition of the "wh-pronoun", the exact words from the instructional statement are retained.

The same keyword deletion or "wh-transformation" rules were used for Rules 3 and 4 with the following rules of paraphrase added, as suggested by Anderson (1972): (a) No substantive terms from the original statement should remain in the paraphrase including all nouns, verbs, and modifiers, and (b) The meaning of the paraphrase sentences(s) should be identical to that in the original. Both the item stem and each foil was paraphrased.

All instructionally-relevant sentences in the prose text used in the experiment were identified and numbered. "Instructionally relevant" sentences

were defined as those which were not simply directions to the student (e.g., "Now, let's examine Step 2"). From the relevant sentences, two random samples were drawn, one for each item writer. An example of one of the sentences is "Cleanability is a requirement of every pontic." This sentence was transformed using Rule 2, for example, into the following test item:

Which is a requirement of every pontic?

*a. cleanability

b. appearance

c. prevention of splinting

d. none of the above

Twelve instructional objectives were also written for the instructional booklet and then given to the item writers for use in producing the objective-based items. Each objective was written so that both a verbatim-recall (VR) or paraphrase-comprehension (PC) item could be produced from it. An example of one of the objectives is:

"Given written descriptions of [verbatim-VR / new-PC] principles, the student will identify which of the three [principles-VR / guiding concepts-PC] is being applied."

Items of each type were assigned randomly to test forms. In each test form, every other item was a rule-generated item, pairs of verbatim items alternated with pairs of paraphrase items, and items written by each experimentor were intermingled to prevent any sequence or fatigue effects due to position of items in the test.

Procedures. Students were given a pretest prior to being given the instructional booklet and then a posttest approximately one month later. A retention test was given ten weeks after the posttest. Students were given different forms of the test at each administration. Forms were assigned to

subjects randomly on the basis of alphabetical groupings of students' names, e.g., all A-D's were given FORM A on the pretest, then FORM B on the posttest. Alphabetical groupings coincided with seat assignments in the dental laboratory classroom where testing was conducted. Groupings numbered N=15, N=18, N=19, and N=20. All test booklets were collected at the end of each testing session. Answer keys and results were given to each student as he completed the posttest and retention test. No feedback was given after the pretest.

Results

The results of this study indicate some similarities and some striking differences between items written by different methods. Item analyses revealed that most items showed "instructional sensitivity" as measured by the difference between pretest and posttest item difficulties. However, 19.8 percent of the items did not show instructional sensitivity. For example, some item difficulties were uniform on pretest and posttest or were lower on the posttest than on the pretest. Importantly, these nonsensitive items were uniformly distributed among the various item types--objective-based, rule-generated, verbatim, paraphrase, and items of different writers.

A further analysis of item difficulties was performed using a 2x2x2x4x3 analysis of variance design in which the factors were: (a) objective-based vs. rule-generated items, (b) verbatim vs. paraphrase items, (c) items of the first writer vs. the second writer, (d) items assigned to each of the four test forms, and (e) a repeated-mesures factor--pretest, posttest, and retention tests. The dependent variable in the analysis was item difficulty. Table 1 summarizes the results of the analysis of variance

- - - - - - - - - - - - -
Insert Table 1 about here
- - - - - - - - - - - - -

A significant difference between rule-generated and objective-based items was found and an inspection of the means shows that rule items were easier overall (mean difficulty .80) than objective-based items (mean difficuly .73). One item writer produced items that were consistently easier (.80), regardless of the techniques used, while the other writer produced consistently more difficult items (.72). There was a difference between the four test forms with mean difficulties of .72, .76, .79, and .80. The expected difference between test occasions was observed with mean difficulty on the pretest of .66, posttest .85, and retention test .79. No difference between verbatim and paraphrase items was observed, nor were there any significant interactions among the factors in the analysis.

In addition to analyses of the differences in mean item difficulties, an examination of the variance of item difficulties was conducted. Variances were computed for each category of item difficulties based on techniques of item writing or test occasion. Also, variances were computed for each of the four rules used to write rule-generated items (and four categories of objective-based items for comparison purposes) to test the notion that item-writing rules provide a unifying and stable influence on item characteristics. Variances for each category of item difficulties are presented in Table 2.

- - - - - - - - - - - - - -
Insert Table 2 about here
- - - - - - - - - - - - - -

As shown in Table 2, rule-generated items showed less variance overall than objective-based items, and this difference is statistically significant, $F(287, 287) = 1.255$, $p < .05$). However, it should be noted that the composite category of item difficulties for Rule 2 across test occasions, which was .0733 as shown in the right-hand column of Table 2, had a higher variance than any of the comparison categories of objective-based items. Hence, variances of all

categories of rule-generated item difficulties were not consistently lower than all categories of objective-based item difficulties.

Rules 1 and 3 which involve the simple deletion of a keyword or phrase were found to have a lower variance (composite variance = .0363) than Rules 2 and 4 (composite variance = .0605) which involves syntactical transformations of instructional sentences in the writing of items. The difference in the variance of these two types of rules was statistically significant, F (153,153) = 1.66, $p < .05$.

As would be expected when instruction is effective, the variance of post-test item difficulties was lower than that of pretest difficulties, with retention test difficulties having a variance in between the other two.

The variance of item difficulties of all verbatim items was .0563, nearly identical to the variance of the difficulties of all paraphrase items which was .0565.

Discussion

Neither the writing of items from instructional objectives nor the use of rules for transforming instructional sentences into test items, as implemented in this study, completely removed subjectivity from item writing. One of two item writers produced consistently more difficult items than the other item writer when both used the same objectives and the same rules. Also, rule-generated items were found to be easier than objective-based items, regardless of the item writer or test occasion.

Since rule-generated items involved verbatim use or paraphrasing of sentences directly from the instructional booklet used in the experiment, it may be that students more easily determined the answers to these items

through syntactic cues or thematic prompts in the sentences or from general test-wiseness. It would not be correct to say that the rule-generated items provided more direct evaluations of instructional effectiveness than objective-based items because of the fact that the rule-generated items were easier on the pretest as well as the posttest.

Rule-generated items were found to have less variability in item difficulties than objective-based items, lending some support to the contention that rules help to produce a more uniform set of items for evaluating learning from instructional materials. However, it is essential that low item variability not be coupled with a lack of instructional sensitivity and this lack was, in fact, found in the present study.

The finding of no significant difference between the item difficulties of verbatim and paraphrased items is suprising and in contradiction to the suggestions of Anderson (1972) and the findings of Bormuth, et al. (1970). It would be expected from these previous findings that verbatim items should be easier than paraphrase items and that paraphrase items should be better evidence for comprehension. Although paraphrasing was done carefully and thoroughly in the present study, it may still be that the difficult task of precise and complete paraphrasing of every word in each item left some undetermined incompleteness. However, the present findings lend some support to the findings of Bortnick (Note 3) in a study of nonsense-syllable learning which provided some evidence for and some evidence against the efficacy of using semantic-substitute questions to assess comprehension.

Conclusion

The present study confirms one of the suggestions of Bormuth (1970, p.10) and Millman (1974, p. 375) that two item writers using an item-writing

approach such as an objective-based one will not produce items of similar quality.

Also, the tentative conclusion of the present study is that the use of rules for transforming instructional sentences into test items will not reduce subjectivity in item writing or produce items of uniform statistical quality if the rules are less than completely operationally defined or automated. If the item wr cer retains functions such as choosing keywords to be deleted from sentences, choosing foils for multiple-choice questions, or syntactically transforming sentences in an nonautomated fashion, the resulting tests will not necessarily be of higher quality than tests written by traditional methods. For example, the present study may indicate as did the study by Richek (Note 4) that the choice of which keyword to delete from an instructional sentence during its transformation will affect the resulting item's difficulty. Richek found that questions eliciting a subject mode were easier than questions eliciting predicate nodes. In the present study, the choice of which keyword to delete was left to the item writer.

The present study suggests that the development of true domain-referenced tests using item forms for evaluating learning from prose materials cannot be done casually, without, perhaps, the somewhat detailed linguistic analyses suggested by Bormuth (1970, Chap. 3) or detailed, linguistic-based algorithms such as the 82-step procedure suggested by Finn (Note 1). It should be cautioned, however, that the present study was not a controlled study comparing automated and nonautomated methods so that this conclusion is somewhat speculative. Clearly there is a need for more empirical study of this important research question.

Table 1

Analysis of Variance on Item Difficulties

Source	df	MS	F
Between Items			
Rule vs. Objective (R)	1	5845.25	5.60*
Verbatim vs. Paraphrase (V)	1	1307.13	1.25
Item Writer (W)	1	8304.06	7.95**
Test Form (F)	3	2980.54	2.85*
RxV	1	181.69	.17
RxW	1	2594.38	2.48
.a :			
TxRxVxWxF	3	2028.13	1.94
Error Between	160	1044.15	
Within Items			
Test Administration (T)	2	16201.88	99.52**
TxR	2	436.77	2.68
TxV	2	47.03	.29
.a :			
TxRxVxWxF	6	37.04	.23
Error Within	320	162.79	

* $p < .05$
** $p < .01$

[a]None of the two-way, three-way, and four-way interactions deleted from this table were significant. The F ratios of the deleted interactions ranged from .13 to 2.68 with associated probabilities above .05.

Table 2

The Variance of Item Difficulties on Three Occasions for Objective-Based and Rule-Generated Items

Type of Item	Test Occasions			
	Pretest	Posttest	Retention	Composite
Rule Generated Items				
Rule 1 - Verbatim	.0568 (24)	.0201 (24)	.0248 (24)	.0362 (72)
Rule 2 - Verbatim	.0906 (24)	.0537 (24)	.0660 (24)	.0733 (72)
Rule 3 - Paraphrase	.0492 (24)	.0246 (24)	.0296 (24)	.0361 (72)
Rule 4 - Paraphrase	.0538 (24)	.0261 (24)	.0489 (24)	.0484 (72)
Composite	.0628 (96)	.0306 (96)	.0426 (96)	.0492 (288)
Objective Items [a]				
Obj. 1-6 - Verbatim	.0719 (24)	.0397 (24)	.0326 (24)	.0540 (72)
Obj. 7-12 - Verbatim	.0719 (24)	.0113 (24)	.0391 (24)	.0571 (72)
Obj. 1-6 - Paraphrase	.0899 (24)	.0396 (24)	.0586 (24)	.0689 (72)
Obj. 7-12 - Paraphrase	.0417 (24)	.0573 (24)	.0498 (24)	.0586 (72)
Composite	.0704 (96)	.0392 (96)	.0462 (96)	.0618 (288)
Composite Totals	.0698 (192)	.0350 (192)	.0450 (192)	.0566 (576)

Note: The number of item difficulties used to compute each variance is shown in parentheses.

[a]Items for objectives 1-6 and those for objectives 7-12 were merged for this analysis so that the number of items used to calculate each variance for each group of items would be uniform and comparable to those for each rule.

REFERENCE NOTES

1. Finn, P. J. *A question writing algorithm.* A paper presented at the meeting of the American Educational Research Association, Washington D. C., March-April, 1975.

2. Hambleton, R. K., Swaminathan, H., Algina, A., & Coulson, D. *Criterion-referenced testing and measurement: A review of technical issues and developments.* Symposium presented at the meeting of the American Educational Research Association, Washington, D. C., March-April, 1975.

3. Bortnick, R. *Effects of a question's form and manner of presentation on its validity.* A paper presented at the meeting of the American Educational Research Association, Washington, D. C., March-April, 1975.

4. Richek, M. A. *The effects of paraphrase alternation and sentence complexity on wh-questions.* A paper presented at the meeting of the American Educational Research Association, Washington D. C., March-April, 1975.

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.

Bormuth, J. R., Manning, J., Carr, J., & Pearson, D. Children's comprehension of between- and within-sentence syntactic structures. *Journal of Educational Psychology*, 1970, *61*, 349-357.

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

Hively, W. Introduction to domain-referenced testing. *Educational Technology*, 1974, *14*, 5-10.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.) *Evaluation in education: current applications*. Berkeley, California: McCutchan Publishing Co., 1974.