DOCUMENT RESUME

ED 190 604 TM 800 396

AUTHOR Roid, Gale: And Others
TITLE A Comparison of Item-Writing Methods for Criterion-Referenced Tests.
INSTITUTION Oregon State System of Higher Education, Monmouth. Teaching Research Div.
SPONS AGENCY Advanced Research Projects Agency (DOD), Washington, D.C.
PUB DATE Apr 80
CONTRACT MDA-903-77-C-0189
NOTE 24p.: Paper presented at the joint Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education (Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests: *Difficulty Level: Elementary Education: Elementary School Teachers: *Item Analysis: Pretests Posttests: Prose: Reading Comprehension: *Reading Tests: Test Construction: *Test Format: *Test Items
IDENTIFIERS *Distractors (Tests)

ABSTRACT

Using informal, objectives-based, or linguistic methods, three elementary school teachers and three experienced item writers developed criterion-referenced pretests-posttests to accompany a prose passage. Item difficulites were tabulated on the responses of 364 elementary students. The informal-subjective method, used by many achievement test developers, allowed maximum freedom of wording and yielded significantly more difficult items. The objectives-based and linguistic methods, in which the item writer chose the foils (distractors), were susceptible to item writer bias. In contrast, the method which provided maximum control, linguistic-based algorithmic foil method, yielded the items which were easy and not subject to bias. It also created items which were insensitive to the pretest-posttest shift in difficulty. Therefore, algorithmic foil methods are promising because they control item writer differences: more research is needed before reasonable item difficulties and instructional sensitivity can be obtained. Teacher-produced items were more sensitive to instruction than those of experienced writers. It is concluded that item-writer differences are real and that field testing is important to identify these differences. (CP)

A COMPARISON OF ITEM-WRITING METHODS FOR CRITERION-REFERENCED TESTS

Gale Roid, Tom Haladyna and Joan Shaughnessy

Teaching Research Division

Oregon State System of Higher Education

Send Correspondence to:

Tom Haladyna

Teaching Research Division

Oregon State System of Higher Education

Monmouth, Oregon 97361

Running Head: A Comparison of Item-Writing Methods

ABSTRACT

Statistical qualities of items were compared across six item writers who used informal, objective and linguistic methods of item writing. Items were written for pretests and posttests to accompany a prose passage. Item difficulties for each type of item were tabulated on the responses of 364 elementary students. Foil-writing methods had significant effects on the pattern of item difficulties of the resulting items. Informal methods of item writing resulted in large differences between item writers. The study indicates the importance of field testing to identify possible item-writer differences in criterion-referenced tests.

A COMPARISON OF ITEM-WRITING METHODS FOR CRITERION-REFERENCED TESTS

A number of measurement theorists (Bormuth, 1970; Hambleton, Swaminathan, Algina & Coulson, 1978; Hively, 1974; Millman, 1974; Popham, 1975) have convincingly argued that criterion-referenced tests should be based on a scientific item-writing technology. This technology begins with domain specification, which is a delineation of the content, subject matter, or job tasks to be tested. For a given domain, a universe or large pool of test items is defined. This universe of test items may be created by an item form (Hively, 1974; Osburn, 1968) or by computerized algorithms (Millman & Outlaw, 1977) or by other methods involving rules or operational definitions of item-writing methods (Bormuth, 1970). A criterion-referenced test is then created by taking a random sample of items from the universe to produce a test form. The score obtained on this random sample of items is a best estimate of a student's performance on the entire domain. Therefore, a criterion-referenced test is defined as assessing "an individual's status [referred to as a domain score] with respect to a well defined behavior domain" (Hambleton et al., 1978, p. 2).

A number of writers, including Anderson (1972), Bormuth (1970) and Millman (1974), have argued that the method of writing tests from written statements of learning objectives does not always include precise domain specification and, therefore, is susceptible to item-writer differences. Roid and Haladyna (1978) demonstrated that two experienced item writers showed significant differences in the difficulty of the items they produced, even though they wrote items using the same learning objectives or loosely-defined rules. Further studies (Roid & Finn, 1978; Roid, Haladyna & Shaughnessy, 1978; Roid, Haladyna, Shaughnessy & Finn, 1978) have further

documented the existence of item-writer differences, as well as the potential for possible control of differences, through algorithmic item-writing techniques.

One emerging technology of criterion-referenced test development has been in the area of item writing for prose learning (Bormuth, 1970). Bormuth described a technology for transforming sentences or elements from prose passages into test questions to assess reading comprehension. This linguistic-based theory of item development has subsequently undergone further research (Finn, 1975; Roid & Finn, 1978; Roid, Haladyna, Shaughnessy & Finn, 1978). There are many learning objectives or instructional systems that require the learning of information through the reading of prose passages or other human verbal communications. The basic elements in these instructional stimuli are sentences which are countable. This collection of countable units can define a domain of content for a course of instruction. Clearly the most precise way to have tests that match the learning objectives and teaching materials is to transform elements of the teaching materials into questions. As Bormuth has described, by transforming sentences from materials, it is possible for the difficulty of the test to be matched exactly to the readability level of the materials. Therefore, a pure form of criterion-referenced tests can be developed for reading comprehension.

The purpose of the present study was to contrast informal, objective-based and linguistic-based methods of writing test items for criterion-referenced tests. The specific research questions of the study were:

1. What differences exist in the characteristics of test items created by various item-writing strategies which ranged from informal-subjective methods to algorithmic methods?

2. To what extent do item-writer differences exist as a function of these various item-writing strategies?

3. Is there a difference in the quality of items produced by experienced item writers as contrasted with teachers for whose students the instructional materials have been selected and assigned?

APPROACH

Subjects

Participants in this study were 364 elementary school students and their teachers from four Oregon public schools. The students ranged considerably in reading ability. Cooperating teachers were asked to select students whose reading ability was sufficient to read the text. In some instances, teachers read the prose passage to some of their students. Students in this study were grouped into a variety of instructional settings (e.g., ungraded or graded). There were no factors present that suggested that this sample of students was significantly different from all students of this age and range of reading levels.

Instruments

Test items were developed from a high interest prose passage in a popular wildlife magazine describing the characteristics of sharks. The reading difficulty level of the passage was approximately fifth grade.

To examine the effects of item-writer variability, six item writers wrote items to assess reading comprehension of the prose passage. Three item writers were researchers who were professionally experienced in item development; the other three writers were elementary school teachers. Each of the six item writers wrote three items for 12 separate test forms, representing different item-writing techniques, which are listed in the left column of Table 1.

On constructing the items for Form 1, the informal-subjective technique, no specifications for writing items were given to the writers. For Forms 2 and 3, six instructional objectives were given to the writers and each writer was asked to construct one item designed to measure achievement of that objective. Both Forms 2 and 3 contained three items from each writer.

Copies of the prose passage were given to a sample of 17 elementary school teachers who were asked to mark with a yellow pen the sentences they believed were most important for their students to learn. Sentences that were chosen by a majority of the 17 teachers were identified. The standard frequency index (Carroll, Davies & Richman, 1971) of each noun and adjective in the chosen sentences was obtained and all nouns or adjectives with a numerical index of 60 or less were identified as high information words, as done in the study by Roid and Finn (1978). Then, 18 rare nouns (high information nouns appearing only once in the passage), 18 key nouns (high information nouns appearing in the passage four or more times), and 18 high information adjectives were identified. Sentences which included these identified words were then used as a basis for item generation for Forms 4 through 12.

For each Form, 4 through 12, each item writer was randomly assigned three sentences of the 18 identified sentences which included the high information words. All writers participated in a training session reviewing the procedures described by Finn (1978) to transform prose sentences into item stems for multiple-choice questions. Using Finn's procedure, each item writer constructed three rare noun, three key noun and three adjective stems.

Then writers prepared the sets of foils for each stem. In the first foil-construction technique, writer's-choice 1 (WC1), writers selected three

single-word foils as possible replacements for the correct answer. In writer's-choice 2 (WC2), the writers were permitted to substitute two or more words as foils for the correct answer. The third foil technique used an algorithm which supplied three individual foil words from the prose passage in the same semantic category as the answer word, using the method of Roid and Finn (1978).

Procedures

Participating teachers volunteered to administer pretests to students identified as able to read the prose passage, provide some instruction, and administer a posttest following this instruction. In one classroom, instruction consisted of simply reading the passage, while in other classrooms, the material was incorporated as part of a larger unit of instruction. Thus, instruction varied considerably for this group of students. No attempt was made to isolate test results by classroom or method of instruction.

Analysis of Data

Student responses to test questions were tabulated and item analyses conducted. An analysis of variance using item difficulty (percentage of students getting the item correct) as the dependent measure was performed. The design for the study was a four-factor analysis of variance with the factors being: (a) 12 item-writing techniques, (b) six item writers, (c) two types of item writers, and (d) the repeated measures of pretest and posttest. The item writers, as a factor, were nested in the third factor, types of item writers.

Interactions involving the repeated measure were examined closely because they suggest that a particular technique or combination of techniques is more or less sensitive to changes in learning that occur as a function of instruction. Therefore, groups of items reflecting a particular item-writing

strategy, item writer, type of item writer (or any combination of these three variables) may show unusually large or small pretest to posttest changes which suggest that a technique or combination of techniques is more or less effective at measuring this change in achievement.

RESULTS AND DISCUSSION

Analysis of Variance Results

The results of the analysis of variance revealed five statistically significant results:

1. Item-writing technique, $F = 3.59$; $df = 11, 144$; $p < .001$.

2. The interaction of item-writing technique with the nested factor, item writers, $F = 1.79$; $df = 44, 144$; $p = .006$.

3. The interaction of item-writing technique with item-writer type, $F = 2.31$; $df = 11, 144$; $p = .012$.

4. The repeated measure, $F = 208.79$; $df = 1, 144$; $p < .001$.

5. The interaction of item-writer type with the repeated measure, $F = 4.02$; $df = 1, 144$; $p = .027$.

All other main effects and interactions were not statistically significant ($p > .05$).

Item-writing techniques. These item-writing techniques ranged from the informal, subjective technique that most achievement test developers have used in the past to several examples of the linguistic-based methods proposed by Bormuth (1970) and refined by Roid and Finn (1978).

Means and standard deviations for these 12 techniques across all conditions of the study appear in Table 1.

Insert Table 1 about here

As recommended by Winer (1962, pp. 77-85) a studentized range statistic was employed to contrast the 12 levels of this independent variable. All pairs of means were contrasted and only four item-writing techniques were found to be significantly different ($p < .05$). The four techniques and their mean difficulties were:

Subjective-informal	41%
Adjective stem with Writer's-Choice 2 foils	44%
Key noun stem with algorithmic foils	57%
Rare noun stem with algorithmic foils	64%

Thus, it seems that the subjective-informal technique leads to significantly more difficult items, while the linguistic-based techniques vary considerably in difficulty as a function of the stem and the foil techniques. What emerges from this analysis is the fact that two of the three item-writing strategies involving algorithm-generated foils were significantly easier than what resulted when using the subjective-informal technique. All three conditions involving the use of the algorithm-generated foils yielded above average difficulties. This is probably due to the fact that these algorithms often introduce obviously incorrect foils so that the student who does not know the correct answer can often deduce the correct answer by eliminating these obviously incorrect foils. These results also indicate that item difficulty is often a function of the method by which items are created, a finding supported in previous studies (Roid & Haladyna, 1978; Roid, Haladyna & Finn, 1978; Roid, Haladyna & Shaughnessy, 1978).

Item-writing technique x item writer. An examination of this result is complicated by the fact that there are 72 cells in this interaction. Winer (1962, p. 232) recommends tests of simple effects to locate the specific sources of significant variation. Unfortunately, tests of simple effects

for each item-writing strategy across the six item writers reveals large and significant F ratios ($p < .001$) for each of the 12 strategies. Consequently, a test of practical significance was employed which is more arbitrary but emphasizes the extent to which certain item-writing strategies interact with item writers to produce excessively hard or easy items. The criterion employed was one involving those entries in Table 1 for the interaction of item-writing strategies and item writers where the means exceeded one standard deviation (16) from the grand mean of 50. Those means observed one standard deviation from the grand mean are underlined in Table 1.

Several findings are suggested from these underlined means. Two out of three of the experienced item writers produced very difficult items using the informal-subjective method of item writing, in contrast to the teachers who produced items of moderate difficulty. Clearly, these demonstrated that an open-ended item-writing technique can have a dramatic effect on item characteristics. Furthermore, the item writer who wrote the most difficult items overall (experienced item writer #2, with a mean of 44 overall), created very difficult items using three techniques involving item-writer freedom--the subjective-informal, the objective-based and writer's-choice version #1 where single rare nouns were provided by the item writer. An intriguing reversal occurred, however, in that the same item writer had very easy items produced with the adjective-stem, algorithmic foil method. This result, coupled with the fact that three other underlined means in Table 1 were for the algorithmic foil method (76, 76 and 72), suggests that the algorithmic foil method affects item writer bias, but does so occasionally by overcorrecting and creating very easy items.

From another perspective, the algorithmic method resulted in extremely easy items in four out of 18 cases (three techniques by six item writers), but the use of algorithms never resulted in items that were extremely difficult. In contrast, when experienced item writers used informal or objective-based techniques, they created very difficult items in three of nine cases, as represented by the nine entries in the upper left of Table 1. Thus, these results reinforce the concept that the idiosyncracies of the item writer, if given a chance to shine through, will have an effect on item difficulties.

Item-writing technique x type of item writer. The preceding interaction dealt with the factor of item writers. The nested factor was type of item writers. The first three item writers were experienced test-item writers, while the last three were teachers. As shown in Table 1, a pattern exists in the interaction of item-writing strategy and type of item writer. Again, a test of simple effects was performed to ascertain under what item-writing strategies the two types of item writers differed. Statistically significant simple effects were detected for three item-writing strategies: (a) informal-subjective, $F = 7.89$, $p = .006$; (b) rare noun/writer's choice 1, $F = 4.55$, $p = .034$; and (c) adjective/algorithm, $F = 7.39$, $p = .007$. For the informal-subjective method, the experienced item writers wrote significantly more difficult items. In the case of the rare noun/writer's-choice 1, a single difficult item by item writer 2 from the experienced writer group was sufficiently low to produce this statistically significant interaction. The third simple effect was a surprising reversal of the tendency of the experienced item writers to write more difficult items. That is, experienced item writers produced significantly easier items when using the adjective-stem technique with algorithmically generated foils. This unique result reinforces the finding mentioned previously that the

algorithmic-foil technique can apparently overcompensate for item-writer effects by creating items that are too easy. A second explanation of this result is that the teachers found the adjective-stem items to be the most challenging of the linguistic-based methods. This was apparently due to the fact that odd changes in wording are sometimes required to transform a sentence into a question when an adjective has been deleted rather than a noun.

Type of item writer and the repeated measure. As reported earlier, the main effect of the repeated measure was highly significant ($F = 208.79$; $df = 1, 144$; $p < .001$). Instruction was effective to the extent that pretest and posttest means differed, 42% on the pretest and 58% on the posttest. This gain of 16%, while statistically significant, is not the kind of gain typically observed in a variety of instructional settings (Haladyna & Roid, 1978). However, interactions of any other independent variable with the repeated measures are one form of evidence that combinations of techniques are more or less effective in measuring achievement that was the target of instruction. In other words, test item groups that obscured this 16% increase in learning are viewed as less reflective of what occurred and, therefore, less sensitive.

The cell and marginal means and standard deviations for the one statistically significant interaction involving the repeated measures are presented in Table 2. As shown there, teachers produced items showing the greatest sensitivity to instruction when compared to experienced item writers. This points out quite dramatically that the type of item writer can affect an important item characteristic, sensitivity to instruction. If this shift can be construed as an indicator of the quality of test items (the tendency to detect real change in performance as a function of learning), then teachers wrote slightly better items than experienced item writers in this experiment.

Insert Table 2 about here

Item Analyses

As a means of further examining factors which potentially underlie the results of this study, item analyses were conducted using an index originally introduced by Cox and Vargas (1966) and recommended by Haladyna and Roid (1978) as useful and appropriate for criterion-referenced tests.

The index used was the pre-to-post difference index (PPDI), the difference in the pretest and posttest difficulties of an item. Those items with PPDI's less than the arbitrary criterion .10 were identified as potentially defective. PPDI's vary from -1.00 to +1.00. A positive PPDI indicates that the item reflects those changes in instruction attributable to learning, while low PPDI's or negative PPDI's may be attributed to either (a) inadequate instruction, or (b) a defective item.

First, items with PPDI's less than .10 were identified, then these items were subjected to inspection to uncover aspects of these items that may have contributed to their low PPDI's as a function of type of item-writing technique, item writer or type of item writer. No items were found to be miskeyed or to contain obvious item-writing defects. Thus, it seems reasonable to conclude that lack of sensitivity for any item was due to either insufficient instruction or a fault in the method used to produce these items. Consequently, the discussion that follows is based on an attempt to uncover systematic trends that can be attributed to peculiarities of item-writing techniques, recognizing that overall the instruction in this experiment was of moderate effectiveness.

PPDI's spanned a very wide range, from -.56 to +.62. Of the 216 items written by six item writers, 63 were found to have PPDI's below .10. An

analysis of these low sensitivity items by item-writing technique revealed that most techniques yielded from three to five insensitive items, including the informal-subjective, objective-based, and linguistic approaches.

Looking at the proportions of insensitive items by item writers, the range was small, from 22% to 39%, with the number of insensitive items ranging from 8 to 14. The factor of item writers was nested in the factor of types of item writers--experienced item writers produced 31% insensitive items when compared to teachers who produced 26%. This finding is in agreement with the earlier report of the interaction of the type of item writer and the repeated measure.

With regard to foil-construction technique, the largest proportions of insensitive items were found with the algorithmic-foil method. This clerical and automated method of assembling foils by selecting words of similar semantic categories from the passage produced the greatest proportions of low sensitivity items found in the study, 44%, 39%, and 50%. This clearly demonstrates that automated foil-generation techniques used with natural language questions are challenging, and will require greater refinement than the technique implemented in the present study.

CONCLUSIONS

The conclusions of the study can be summarized in relation to the three main research questions of the study dealing with differences between item-writing techniques, item-writer bias, and types of item writers.

Item-writing techniques. The technique which provided the maximum amount of freedom of wording to the item writer was the informal-subjective method in which no objectives or rules of item writing were used. This method resulted in significantly more difficult items, and was susceptible to large differences between item writers. Also, objective-based and

linguistic methods in which the item writer chose the foils were found to be susceptible in some instances to item-writer bias. In contrast, a reversal was found with the techniques which provide the maximum amount of control of item-writer differences--the linguistic-based, algorithmic-foil method. In this method, sentences from the prose passage are transformed into questions by the item writer, and then foils are added clerically by taking them from word lists created from the passage. This method resulted in the easiest items, and created the highest proportion of items that were insensitive to the pretest-posttest shift in difficulty. These findings suggest that some degree of item-writer choice in wording may be necessary in the context of writing items for reading comprehension, until foil-writing algorithms can be developed that are more sensitive than those used in the present study.

Perhaps the technique that survived the comparisons of the study with the fewest limitations was the linguistic-based method involving the use of nouns and the second writer's-choice method of foil construction. In this technique, item writers transformed sentences into questions by deleting a noun phrase and then forming foils by inserting a noun phrase in place of the one deleted. This method resulted in good instructional sensitivity (average of .20 difference between pretest and posttest item difficulties) and a relatively homogeneous level of item difficulties across item writers.

Item-writer differences. The finding that algorithmic-foil construction, as implemented in this study, leads to items that are very easy is similar to that found in earlier studies by Roid and Finn (1978) and Roid, Haladyna, Shaughnessy and Finn (1978). The one advantage of algorithmic

methods of foil construction, which could not be examined in the present study, but which was established in Roid, Haladyna, Shaughnessy and Finn (1978), is that they control the variability of item difficulties between item writers. Therefore, algorithmic foil methods may be promising in that they control item-writer differences, but need to be designed so that reasonable item difficulties and instructional sensitivity can be obtained. Further research and development in foil-construction techniques is clearly needed.

Types of item writers. It would be tempting but, of course, imprudent to conclude from this study that the better item writers were teachers, who may know their students more personally than a group of experienced item writers. The more important meaning of the finding that item writers of different backgrounds differed significantly in the resulting instructional sensitivity of their items is that item-writer effects are real and significant. The challenge is to develop tests with careful specification of the item-writing methods to be used, so that these biases can be identified and isolated. The only way these effects can be identified is through field testing of items on students. Thus, the present study can be seen as providing evidence for the importance of empirical item review. Not that items are to be selected or discarded from a domain on the basis of their statistical qualities, but rather that the item specifications be changed if faulty items result. Documented item-writing methods can lead to criterion-referenced tests that have the desirable property of being random samples of items from a well-specified domain (Hambleton et al., 1978, p. 38). Field testing can then provide quality control through the assessment of the statistical qualities of items written by different people and different methods.

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.

CARROLL, J. B., DAVIES, P., & RICHMAN, B. Word frequency book. Boston: Houghton-Mifflin, 1971.

COX, R. C., & VARGAS, J. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1966.

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

FINN, P. J. Generating domain-referenced, multiple-choice test items from prose passages. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

HALADYNA, T., & ROID, G. The role of instructional sensitivity in the empirical review of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

HAMBLETON, R. K., SWAMINATHAN, H., ALGINA, J., & COULSON, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research. 1978, 48, 1-47.

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 Company, 1974.

MILLMAN, J., & OUTLAW, W. S. Testing by computer. Ithaca, New York: Cornell University Extension Publications, 1977.

OSBURN, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104

POPHAM, W. J. Educational evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

ROID, G. H., & FINN, P. J. Algorithms for developing test questions from sentences in instructional materials (NPRDC Tech. Tep. 78-23). San Diego: Navy Personnel Research and Development Center, 1978.

ROID, G. H., & HALADYNA, T. M. A comparison of objective-based and modified-Bormuth item writing techniques. Educational and Psychological Measurement, 1978, 35, 19-28.

ROID, G., HALADYNA, T., & FINN, P. A comparison of several multiple-choice, linguistic-based item writing algorithms. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

ROID, G., HALADYNA, T., & SHAUGHNESSY, J. A comparative study of informal objective-based and linguistic-based item-writing methods. Monmouth, Oregon: Teaching Research, 1978.

ROID, G., HALADYNA, T., SHAUGHNESSY, J., & FINN, P. Item writing for domain-based tests of prose learning. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

WINER, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

FOOTNOTE

This research was supported by contract number MDA-903-77-C-0189 from the Defense Advanced Research Projects Agency. Dr. Pat-Anthony Federico of the Navy Personnel Research and Development Center, San Diego, was the technical monitor for this project.

Views expressed in this article are those of the authors and not necessarily those of the supporting agencies.

Acknowledgments are extended to Dr. John Bormuth of the University of Chicago, Dr. Patrick Finn of the State University of New York at Buffalo, and Dr. Jason Millman of Cornell University, who contributed to aspects of this research. Appreciation is also extended to Dr. Harold F. O'Neil, Jr., of the Army Research Institute, for his encouragement in the formative stages of this research.

Table 1

Cell and Marginal Means and Standard Deviations of Item Difficulties[1] for the Main Effects of Item Writing Techniques (A), Item Writers (B), Types of Item Writers (C), and Interactions Between A and B, and B and C

Item-Writing Techniques	Experienced Item Writers								Teachers								TOTAL	
	1		2		3		Total		1		2		3		Total			
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
1. Informal-Subjective	26	13	34	2	37	5	33	9	57	14	52	18	42	7	51	13	41	14
2. Objective 1	38	4	45	12	64	20	48	17	58	9	47	21	42	15	49	15	49	16
3. Objective 2	50	20	30	7	58	10	46	17	49	8	36	13	63	9	49	15	48	16
4. Rare Noun/WC1	58	10	26	23	45	18	43	21	42	8	75	7	53	8	56	16	50	19
5. Rare Noun/WC2	55	8	48	22	43	10	49	14	45	11	71	8	56	10	58	14	53	14
6. Rare Noun/Algorithm	76	6	60	16	52	17	62	16	59	9	76	4	59	11	65	12	64	14
7. Key Noun/WC1	58	8	40	11	41	10	46	12	49	15	61	24	50	7	53	16	50	14
8. Key Noun/WC2	58	5	39	18	39	11	45	14	40	9	54	9	56	2	50	10	48	12
9. Key Noun/Algorithm	64	15	46	15	54	3	56	13	62	12	45	5	72	8	60	14	57	13
10. Adjective/WC1	46	18	53	18	56	6	51	14	48	28	41	11	36	10	41	17	46	16
11. Adjective/WC2	42	15	36	28	57	10	45	19	45	16	39	11	42	10	42	11	44	15
12. Adject./Algorithm	55	20	69	23	59	11	61	17	50	14	44	10	38	17	44	13	52	17
TOTAL	52	17	44	19	50	13	49	17	50	13	53	18	51	14	52	15	50	16

[1]Entries are means and standard deviations of average item difficulties as calculated across both pretests and posttests.

Table 2

Means and Standard Deviations

for the Interaction of the Repeated Measure with Type of Item Writer

	Pretest		Posttest		Total	
	M	SD	M	SD	M	SD
Experienced Item Writers	42	19	56	17	49	19
Teachers	42	19	61	16	51	20
TOTAL	42	19	58	17	50	16

AUTHORS

ROID, GALE. Address: Horizon Associates, 763 Caroline Way N., Monmouth, Oregon, 97361. Title: Consultant. Degrees: A.B. Harvard University, M.A., Ph.D. University of Oregon. Specialization: Measurement; Evaluation; Instructional Improvement.

HALADYNA, TOM Address: Teaching Research Division, Oregon State System of Higher Education, Monmouth, Oregon, 97361. Title: Research Professor. Degrees: B.A. Illinois State University, M.A. San Jose State University, Ph.D. Arizona State University. Specialization: Educational Measurement; Statistics; Research Methodology.

SHAUGHNESSY, JOAN. Address: Teaching Research Division, Oregon State System of Higher Education, Monmouth, Oregon, 97361. Title: Instructor. Degrees: B.A. Indiana University, M.A. Wichita State University, Ph.D. candidate Michigan State University.. Specialization: Child Development.