DOCUMENT RESUME

ED 085 401 TM 003 342

AUTHOR Pyrczak, Fred
TITLE Subjective Evaluation of the Quality of Standardized Reading - Comprehension Items.
SPONS AGENCY National Science Foundation, Washington, D.C.
NOTE 11p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Content Analysis; Item Analysis; *Reading Comprehension; *Reading Tests; *Standardized Tests; *Test Reviews

ABSTRACT

A random group of 49 items was drawn from nine commerically available reading comprehension tests. Each test was classified independently by two judges as either a measure of the ability to find answers to questions answered explicitly or in paraphrase in the passages, a measure of the ability to draw inferences or deductions, or a measure of some "other" skill. Both judges classified a majority of the items as measures of the ability to draw inferences or deductions, and there was a reasonable amount of agreement between the judges in this classification process. The judges also indicated the extent to which they thought seven types of faults were present in each item. One judge found a total of 122 faults in the 49 items: the other judge found 31. The judges were most often in agreement in judging items to be measures of general knowledge rather than measures of the ability to comprehend specific passages. (Author)

ABSTRACT

Subjective Evaluation of the Quality of Standardized Reading-Comprehension Items

Fred Pyrczak
California State University, Los Angeles

A random group of 49 items was drawn from nine commercially available reading comprehension tests. Each test was classified independently by two judges as either a measure of the ability to find answers to questions answered explicitly or in paraphrase in the passages, a measure of the ability to draw inferences or deductions, or a measure of some "other" skill. Both judges classified a majority of the items as measures of the ability to draw inferences or deductions, and there was a reasonable amount of agreement between the judges in this classification process. The judges also indicated the extent to which they thought seven types of faults were present in each item. One judge found a total of 122 faults in the 49 items; the other judge found 31. The judges were most often in agreement in judging items to be measures of general knowledge rather than measures of the ability to comprehend specific passages.

Subjective Evaluation of the Quality of Standardized Reading-Comprehension Test Items[1]

Fred Pyrczak
California State University, Los Angeles

Writing multiple-choice items of high quality requires considerable insight into the content and intellectual skills that are to be measured, the desirable characteristics and limitations of multiple-choice items, and the probable reactions of examinees to the items. Because item-writing is a complicated skill, it is not surprising that a relatively large number of faulty items in standardized tests have been identified by subject-matter specialists and scholars (e.g., Hoffmann, 1962). The basic purpose of the present study was to determine the extent to which faults are present in the items in a specific set of standardized reading-comprehension tests. The subjective analysis of item quality conducted in this study differed from earlier analyses in three important respects: a sample of items was drawn systematically for analysis in this study, two judges independently rated the quality of each item, and both judges used the same rating scale when evaluating each item.

PROCEDURES

Sample. A set of nine standardized reading-comprehension tests, which are listed in the Test Reference List at the end of this paper, were selected for use in this study. All of the tests were currently available

[1]This study was supported by a NSF Institutional Grant awarded by the Faculty Awards Committee of the California State University, Los Angeles.

from commercial publishers at the time of the study, and all were designed for use with junior- and senior-high school students. Only those items that ask questions about specific reading passages presented in the tests were used. Because in most of the tests more than one question is asked about each passage and because it was desirable to examine the possible interrelatedness of the items for a given passage, a sample of items was drawn indirectly by random selection of passages from each test. Passages were selected randomly from each test until at least five per cent of the total number of items were included in the sample. No more than ten per cent of the items from any given test were included in the total sample of items. A total of 49 items was selected.

Analysis. Each item selected for use in this study was evaluated independently by two judges.[2] A special rating form was developed to aid the judges. The first part of the form asked the judges to indicate the skill they thought each item was designed to measure: (1) finding the answers to questions answered explicitly or in paraphrase, (2) drawing inferences or deductions, or (3) some "other" skill.

The second part of the form presented the judges with seven potential item faults. These were:

1. Inadequate keyed choice (i.e., the choice designated as "correct" is not thoroughly correct).
2. Defensible distracter(s) (i.e., one or more "incorrect" choices can be defended as the correct choice).

[2]William R. Crawford, University of California, Los Angeles and Mary B. Willis, American Institutes for Research, Palo Alto served as the judges.

3. Information other than that provided in the passage is needed in order to identify the keyed choice.

4. Question measures general knowledge (i.e., examinees may be able to answer on the basis of their knowledge without reading the associated passage).

5. Item is related to another item on the same passage in such a way that the interrelationship may aid an examinee who has not carefully considered the passage.

6. Distracters are not homogeneous with keyed choice (i.e., keyed choice is more general, longer, etc.).

7. Other faults.

Faults three, four, and five refer specifically to multiple-choice items designed to be passage-dependent. These faults have been discussed at length by Pyrczak (1972, 1973a).

For each item, the judges were asked to indicate which faults, if any, were present. For each fault, furthermore, the judges were asked to indicate the extent to which the fault is detrimental to the item's ability to discriminate between those who do and those who do not have the reading skill in question by checking either "not detrimental" "moderately detrimental," or "seriously detrimental." A similar three-point rating scale previously has been used successfully in evaluating the quality of arithmetic-reasoning items (Pyrczak, in press). The judges also were asked to give a written explanation for each fault that they found.

RESULTS

Skills measured. One judge classified 14 of the items as measures of the ability to find answers to questions answered explicitly or in paraphrase in the passages, 31 as measures of the ability to draw inferences

or deductions, and 4 as measures of some "other" skill. The other judge classified 20, 26, and 2 items as belonging to these three skill areas, respectively. The second judge did not indicate the type of skill measured by one of the items. The two judges agreed on the classification of 31 of the 49 items. This is a fairly high rate of agreement considering the types of judgments involved. Pyrczak (1973b) has discussed some of the problems involved in classifying reading-comprehension items in terms of the skills they appear to measure.

Faults present. One judge found a total of 122 faults in the 49 items while the other judge found only 31 faults. Clearly, the two judges applied different standards when rating the items and had different types of insights into the content of the items and their relationships with the passages. Thus, by conventional standards there was a low rate of interobserver agreement. Table 1 indicates the number of times both judges agreed that a particular type of fault was present in a given item. It is interesting to note that both judges thought that seven items, to some extent, were measures of general knowledge.

INSERT TABLE 1 ABOUT HERE

Table 2 shows the number of faults found in the 49 items by each judge. It is interesting that each judge found each type of fault at least once.

INSERT TABLE 2 ABOUT HERE

DISCUSSION

A major weakness of the present study was the low rate of agreement between the judges on the presence or absence of faults in the items. While the rate of agreement was disappointingly low, it was not especially surprising considering the subtle factors involved in the types of judgments that the experts were asked to make. It is interesting to note that as part of a larger study Pyrczak (in press) had arithmetic-reasoning items rated for quality by three judges using a check list similar to that used in this study and obtained fairly consistent ratings. Thus, it may be that making judgments of the quality of arithmetic items is a more clear-cut process than making judgments of the quality of reading-comprehension items. Clearly, further investigation is needed to determine if procedures can be developed for obtaining consistent, independent judgments of the quality of reading-comprehension items. Such procedures would be very helpful when editing and revising reading-comprehension items during test construction.

Because of the limitations of the rating process, it is difficult to draw an overall generalization regarding the extent to which faults are present in standardized reading-comprehension tests. It seems reasonable to conclude, however, that a majority of the items will be subject to some type of criticism if carefully examined by experts.

The judges most often agreed on the absence of passage-dependence due to items measuring general knowledge as a fault. Pyrczak (1972) suggested an empirical method of identifying items with this fault. Specifically, he suggested administering reading-comprehension questions

in the absence of the associated passages and asking examinees to indicate the basis or bases for their responses.

In conclusion, a majority of the items in reading-comprehension tests appear to be measures of the ability to draw inferences and deductions from reading material, and a majority appear to be subject to some type of criticism when critically examined by experts. Obtaining agreement among experts on the number and nature of the faults in a given item when it is examined independently by them appears to be difficult and is a topic that deserves further investigation.

BIBLIOGRAPHY

Hoffman, B. The tyranny of testing. New York: Crowell-Collier Press, 1962.

Pyrczak, F. Objective evaluation of the quality of multiple-choice items designed to measure reading-comprehension. Reading Research Quarterly, Fall, 1972, 8, 62-71.

Pyrczak, F. Special factors to consider when selecting reading-comprehension tests and exercises. Reading Improvement, Spring 1973a, 10, 37-38.

Pyrczak, F. Subjective analysis of the skills measured by selected reading tests designed for use in high school. Unpublished paper, 1973b.

Pyrczak, F. Validity of the discrimination index as a measure of item quality. Journal of Educational Measurement (in press).

TEST REFERENCE LIST

California Achievement Tests: Reading, Level 5, Form A. Monterey: CTB/ McGraw-Hill, 1970.

Comprehensive Tests of Basic Skills: Reading Comprehension, Level 4, Form Q. Monterey: McGraw-Hill, 1968.

Cooperative English Tests: Reading Comprehension, Form 2A. Princeton: Educational Testing Service, 1960.

Davis Reading Test, Form 1A. New York: Psychological Corporation, 1956-1957.

The Nelson-Denny Reading Test: Comprehension Test, Form A. New York: Houghton Mifflin, 1960.

Stanford Achievement Test: High School Reading Test, Form W. New York: Harcourt, Brace and World, 1965.

Sequential Tests of Educational Progress: Reading, Series II, Form 2A. Princeton: Educational Testing Service, 1969.

Tests of Academic Progress, Form S. New York: Houghton Mifflin, 1971.

Traxler High School Reading Test, Form A. Indianapolis: Bobbs-Merrill, 1966.

Table 1: Number of times both judges found each fault in a given item.

Fault	Number of times both judges found the fault in an item
Inadequate keyed choice	0
Defensible distracter(s)	3
Information other than that provided in the passage is needed in order to identify the keyed choice	2
Question measures general knowledge	7
Item is related to another item on the same passage in such a way that the interrelationship may aid an examinee who has not carefully considered the passage	0
Distracters are not homogeneous with keyed choice	2
Other faults	3

Table 2: Number of faults found in the 45 items by each judge.

Fault	Judge 1	Judge 2
Inadequate keyed choice	18	1
Not detrimental	9	0
Moderately detrimental	5	0
Seriously detrimental	4	1
Defensible distracter(s)	25	5
Not detrimental	6	0
Moderately detrimental	10	2
Seriously detrimental	9	3
Information other than that provided in the passage is needed in order to identify the keyed choice	16	5
Not detrimental	10	1
Moderately detrimental	4	0
Seriously detrimental	2	4
Question measures general knowledge	22	7
Not detrimental	6	2
Moderately detrimental	13	3
Seriously detrimental	3	2
Item is related to another item on the same passage in such a way that the interrelationship may aid an examinee who has not carefully considered the passage	10	1
Not detrimental	2	0
Moderately detrimental	7	0
Seriously detrimental	1	1
Distracters are not homogeneous with keyed choice	12	4
Not detrimental	7	0
Moderately detrimental	4	2
Seriously detrimental	1	2
Other faults	19	8
Not detrimental	10	1
Moderately detrimental	8	1
Seriously detrimental	1	6
Total	122	31