ED 208 028 TM 810 723

AUTHOR Oosterhof, Albert C.; Coats, Pamela K.

TITLE Comparison of Difficulties and Reliabilities of Math-Completion and Multiple-Choice Item Formats.

INSTITUTION Florida State Univ., Tallahassee.

PUB DATE [81]

NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Comparative Analysis; *Difficulty Level; Grading; Higher Education; *Mathematics; *Multiple Choice Tests; Quantitative Tests; *Test Format; Testing Problems; Test Items; *Test Reliability; Test Validity

IDENTIFIERS *Short Answer Tests; Test Length

ABSTRACT

Instructors who develop classroom examinations that require students to provide a numerical response to a mathematical problem are often very concerned about the appropriateness of the multiple-choice format. The present study augments previous research relevant to this concern by comparing the difficulty and reliability of multiple-choice and completion item formats as applied to the classroom measurement of quantitative skills. This investigation also includes two variations of the multiple-choice format designed to reduce cues provided by alternatives. Focus is placed on the external validity of the experiment by using an actual examination of course material administered to students in a realistic classroom setting. When plausible distractors are used, minimal effects on difficulty and reliability are observed as a result of using "none of the above" or by using ranges of values for alternatives. The results of the study also support serious consideration of the math-completion format when efficiency of scoring is not a major concern. It is shown that fewer math-completion items are required for obtaining reliability equal to that provided by multiple-choice items. Implications which varying difficulties and reliabilities have on grading standards and test length are discussed. (Author/AL)

COMPARISON OF DIFFICULTIES AND RELIABILITIES OF MATH-COMPLETION AND MULTIPLE-CHOICE ITEM FORMATS

Albert C. Oosterhof and Pamela K. Coats

Florida State University

Authors of educational measurement texts generally favor use of test items which require making a choice among specified alternatives in contrast to items which require the examinee to produce a limited free response. Wesman (1971) recommends against the use of short-answer items concluding their superiority over selection-type items is more apparent than real in actual testing situations. Ebel (1979) indicates that short-answer items are used mainly to test for factual information, and that good objective test items do not permit identification of the correct response on the basis of simple recognition or sheer rote memory. Popham (1981) takes a more cautious approach by suggesting a major weakness of multiple-choice items is the ability of examinees to recognize correct answers that, without assistance, they would not be able to construct.

Instructors who develop classroom examinations that require students to provide a numerical response to a mathematical problem are often very concerned about the appropriateness of the multiple-choice format. The present study augments previous research relevant to this concern by comparing the difficulty and reliability of multiple-choice and completion item formats as applied to the classroom measurement of quantitative skills. This investigation also includes two variations of the multiple-choice format designed to reduce cues provided by alternatives. Focus is placed on the external validity of the experiment by using an actual examination of course material administered to students in a realistic classroom setting. Implications which varying difficulties and reliabilities have on grading standards and test length are discussed.

Background

The literature contains a limited number of investigations comparing math-completion and various multiple-choice formats. Wesman and Bennett (1946) used a multiple-choice test battery administered to nursing school applicants. A portion of subjects were administered a modified form of the test in which the fifth alternative was changed to "none of these." The difficulty and item-test correlations of test items that measured arithmetic skills were on the average quite similar for the versions.

Frederickson and Satter (1953) discussed the development of the Navy Arithmetical Computation Test and demonstrated the appropriateness of constructing multiple-choice alternatives from answers generated from completion items. Shifts in item difficulty from the free-answer to the multiple-choice forms were found to be relatively small. Rimland and Zwerski (1962) reported similar findings in the development of the Navy Arithmetic Test.

Traub and Fisher (1977) compared the equivalence of constructed-response and multiple-choice formats on mathematical reasoning and verbal comprehension subtests. Eighth-grade students were initially administered items in the constructed-response format. To control for the retention effect inherent in a study by Heim and Watts (1967) using verbal items, Traub and Fisher administered items rewritten in the multiple-choice format two weeks later. Mean test scores were 3% to 6% lower when items were written in the multiple-choice format. *Alpha* reliability coefficients for alternate forms of the 30-item math test were, with one exception, between .84 and .87. Using a procedure suggested by Lord (1971) for assessing equivalence, the tests of mathematical reasoning were found to measure the same psychological dimensions independent of item format. Approximately nine hours was required in the Traub and Fisher study to administer the battery of instruments. Student motivation was recognized as a problem within the experimental conditions.

The present investigation evaluated math-completion and selected multiple-choice item formats for equivalence in difficulty and reliability when administered under conditions representative of classroom examinations. Alternate item formats were administered concurrently to groups of examinees equated through random assignment. Multiple-choice options were formulated by the instructor using experiential knowledge of common errors instead of from responses empirically derived from previous free response forms of the item. "None of the above" and ranges of numerical responses were investigated as possible techniques for reducing the effect providing the student with response options may have on identifying the correct answer.

Method

An examination in a business finance course was used in the investigation. The examination was developed by the instructor using test development and item construction principles discussed in most introductory measurement texts. The test length varied from 34 to 40 items across the academic terms in which the study was conducted.

Skills assessed by 12 test items were identified for use in the study. Each of the 12 items was written in the following four formats (abbreviated identifiers are given in parentheses):

1. *Completion*.

2. Multiple-choice using a single numerical value for each of five alternatives; each of the distractors represented common errors (*5-Values*).

3. Multiple-choice as above, except the fifth alternative was replaced with "none of the above" (*N of Above*).

4. Multiple-choice using ranges of values incorporating all possible values of the examinee's answer; ranges of each alternative respectively encompassed the five numerical values used above (*Ranges*).

A common stem was used across the four forms of each test item. The Figure illustrates how an item was adapted to each of the formats.

Insert Figure about here

Four forms of the examination were prepared. Table 1 describes how the 12 items included in the investigation appeared in the same order within each form, but in different formats across the four forms. Each triad of items used an A, C, or E as the correct multiple-choice alternative, but not necessarily in that order. The 12 items were administered to undergraduate business majors as part of a course examination in each of three academic terms. The four forms were randomly ordered before being distributed to students each term. The total number of students assigned to each of the forms is indicated in Table 1.

Insert Table 1 about here

All forms of the test shared a common scoring key with the exception of items written in the completion format. Responses were recorded by examinees on machine readable answer forms except that answers to the completion items were initially recorded in the test booklets. The instructor scored responses to the completion items and marked the keyed response (A, C, or E) on the student's answer form if the response was found to be correct. The answer forms were then machine scored with all items scored dichotomously.

Item p-values were calculated separately for the 12 items written in each of the four formats. The weighted mean difficulty was then established for each item format. Items incorporating "none of the above" as a response alternative were further analyzed by comparing the difference in item difficulty that occurred as a function of whether this alternative represented the correct response.

To facilitate discussion of the findings, four expanded tests were conjectured, each consisting of 40 items equivalent to the completion of one of the three multiple-choice type of items included in the present investigation. Setting item difficulty, variance, and covariance consistent with those observed in the study, means and standard deviations of scores on the conjectured tests were estimated. Assuming a fixed shape to the distribution of scores, percentile ranks associated with specific criterion scores were also estimated for each of the four expanded tests.

The KR-20 reliability coefficient was calculated for each triad of items within each of the four item formats, and a pooled estimate obtained for each format. The Spearman-Brown formula was used to calculate reliabilities for 40-item tests consisting of equivalent items. The formula was also used to determine the ratio of items required for reliability equal to that of the completion item format.

Results

Observed p-values for the 12 items within each of the four formats are listed in Table 2. The items incorporated in the investigation are mostly

Insert Table 2 about here

of moderate difficulty with the middle 50% of the values ranging between .475 and .695. Even with a somewhat restricted range of difficulties, correlations between rankings of p-values ranged from .72 to .91. Completion and 5-Values had the highest correlations with alternate formats, whereas N of Above had the lowest.

Completion items were consistently the most difficult, with the three multiple-choice formats being of near-equal difficulty. Providing ranges of values for alternatives in contrast to specific numerical values did not affect item difficulty overall. Table 3 illustrates how substituting "none of the above" as an option generally made the item more difficult, almost all the increased difficulty occurring when "none of the above" was the correct answer.

Insert Table 3 about here

Table 4 presents the means and standard deviations that were projected for a 40-item test. Assuming normal distributions of scores for each of the tests (a condition that in reality may not be true), percentile equivalents across the four formats can be established as illustrated in Table 5. Scores which were equivalent to selected percentile ranks for Completion items were computed first, and the percentile ranks of these scores for each of the multiple-choice formats subsequently determined. For example, a projected score of 16.918 would represent the 40th percentile for Completion items, but only 20%, 23%, and 19% of the examinees would be expected to score below this score when administered corresponding tests using the respective multiple-choice item formats.

Insert Table 4 about here

Insert Table 5 about here

The pooled estimate of reliabilities associated with the four item formats is presented in Table 6. Estimates of reliability based on triads of items and

Insert Table 6 about here

then pooled across the four forms of the test suggest a discrepancy between Completion and the multiple-choice formats. Among the three multiple-choice formats, 5-Values resulted in the highest reliability and Ranges in the lowest. When adjusted to 40-item tests with the Spearman-Brown formula, all formats resulced in high reliabilities. However, Table 6 also indicates that a significant proportion of additional multiple-choice items would be required to obtain reliability equivalent to the Completion format. For example, it is estimated that 62, 70, and 73 items of the respective multiple-choice formats would be required to match the reliability of 40 Completion items.

Discussion

Differences in item difficulty are most significant between Completion and each of the multiple-choice formats. Mean difficulties for the respective formats suggest that providing examinees with alternative answers results in test scores approximately 20% to 30% higher than when a math-completion format is used. (This is inconsistent with the findings of other research studies referenced previously.) It is probable that examinees rework a problem presented in the 5-Value format if the worked solution is inconsistent with all five alternatives. If a solution consistent with an alternative can not be obtained, the examinee will likely choose the alternative perceived most consistent with the obtained solution to the problem. Only if the foils are able to encompass a high proportion of incorrect solutions or the correct solution is perceptually deviant from probable incorrect solutions in a manner not discernible to test-wise behavior would a 5-Value format not provide the examinee with cues to the correct answer.

The substitution of "none of the above" for the fifth alternative appears to have an insignificant effect on item difficulty unless it is the correct response. Possibly examinees are leery of using this alternative unless they are confident of their solution. Indeed, on an average, the difficulty of the N of Above format is very similar to that observed with the Completion format when "none of the above" is the correct response. To suggest that "none of the above" be used perpetually as the correct alternative is tempting.

The Ranges and 5-Values formats resulted in equivalent overall item difficulties. Ranges does not provide the same degree of feedback to incorrect solutions as does 5-Values, but may permit selection of the keyed response by obtaining a nearly correct solution for the wrong reason. Ranges will also probably promote caution when an examinee's solution deviates dramatically from the ranges of values used for alternatives. Increasing ranges of values associated with each alternative would reduce the latter problem with a consequential increase in the former.

Estimates obtained from the present study suggest that a distribution of test scores will vary noticeably as a function of the item format used. Table 5 indicates how the greatest differences would be expected between Completion items and the various multiple-choice formats. Distributions of scores may not be normal as was assumed for calculating percentiles, however differences in means and variability of test scores resulting from varying item formats probably is sufficiently significant to merit reestablishing standards if meaningful changes are made in the portions of math completion and multiple-choice items included in tests.

The reliability of all four item formats is respectable. However, the higher reliability of the Completion items implies that approximately 50% to 80% additional multiple-choice items are required to obtain reliability equivalence to the math-completion format. The instructor may wish to determine the point at which creation of effective response foils, generation of additional items, and subsequent need for more time in the classroom to administer longer tests are compensated by the greater scoring efficiency of multiple-choice items.

The authors find minimal advantage, when using a multiple-choice format, to camouflage the correct response by using either a "none of the above" response or by using ranges of numerical values for each alternative. The results of the study also support serious consideration of the math-completion format when efficiency of scoring is not a major concern. Generalization from this research context to other measurement settings must be done cautiously. Subjects included in the present study were fairly competitive college students who were being assessed on relatively complex mathematical problems. If for example the investigation were replicated with less motivated students, selection of a multiple-choice alternative may be more a function of guessing as was the case in the Traub-Fisher study. More frequent guessing might reduce further the lower reliability of multiple-choice items.

References

Ebel, R. L. Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc., 1979.

Frederiksen, N., and Satter, G. A. The construction and validation of an arithmetic computation test. Educational and Psychological Measurement, 1953, 13, 209-227.

Heim, A. W., and Watts, K. P. An experiment on multiple-choice versus open-ended answering in a vocabulary test. British Journal of Educational Psychology, 1967, 37, 339-346.

Lord, F. M. Testing if two measuring procedures measure the same psychological dimension. (Research Bulletin RB-71-36), Princeton, NJ: Educational Testing Service, 1971.

Popham, W. J. Modern educational measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1981.

Rimland, B., and Zwerski, E. The use of open-end data as an aid in writing multiple-choice distracters: An evaluation with arithmetic reasoning and computation items. Journal of Applied Psychology, 1962, 46, 31-33.

Traub, R. E., and Fisher, C. W. On the equivalence of constructed-response and multiple-choice tests. Applied Psychological Measurement, 1977, 1, 355-369.

Wesman, A. G. Writing the test item. In Educational Measurement (2nd ed.), R. L. Thorndike (Ed.), Washington, D.C.: American Council on Education, 1971.

Wesman, A. G., and Bennett, G. K. The use of 'none of these' as an option in test construction. Journal of Educational Psychology, 1946, 37, 541-554.

Item Stem

If Internal Rate of Return equals 11 percent, Profitability Index equals 1, and the Present Value of the after-tax cash flows over the life of the project equals $268.13, what is the initial cash outlay?

Response Variations

<u>Completion:</u>

ANSWER ________________

<u>5-Values:</u>

A. $268.13
B. $294.00
C. $313.07
D. $326.00
E. $358.00

<u>N of Above:</u>

A. $268.13
B. $294.00
C. $313.07
D. $326.00
E. None of the above

<u>Ranges:</u>

A. Less than $275
B. Between $275 and $300
C. Between $300 and $325
D. Between $325 and $350
E. Greater than $350

Figure. Illustration of an item adapted to the four formats.

TABLE I

Format of Items and Number of Subjects Assigned to Each Form

		Form of Test			
Item	Key	A	B	C	D
1	C	Completion	5-Values	N of Above	Ranges
2	A				
3	E				
4	A	Ranges	Completion	5-Values	N of Above
5	E				
6	C				
7	E	N of Above	Ranges	Completion	5-Values
8	A				
9	C				
10	C	5-Values	N of Above	Ranges	Completion
11	E				
12	A				
Number of examinees administered each form		60	59	57	56

TABLE 2

Item Difficulties Listed
by Item Format

Item	Completion	5-Values	N of Above	Ranges
1	.367	.559	.509	.583
2	.483	.661	.737	.542
3	.250	.441	.368	.500
4	.627	.825	.792	.817
5	.644	.860	.645	.717
6	.695	.789	.667	.750
7	.404	.500	.317	.475
8	.439	.541	.633	.695
9	.702	.875	.283	.831
10	.542	.550	.678	.543
11	.562	.600	.729	.667
12	.188	.300	.119	.368
Average	.492	.623	.589	.626

TABLE 3

Differences in p-Values Between N of Above and Other Item Formats

	Difference from Completion	Difference from 5-Values	Difference from Ranges
Average differences for all 12 items	.098	-.035	-.034
Average differences for 4 items keyed E	.050	-.086	-.075
Average differences for 8 items not keyed E	.122	-.010	-.014

Negative value indicates that item presented in N of Above format was more difficult than when presented in alternate format.

TABLE 4

Projected Means and Standard Deviations of 40-Item Tests

	Completion	5-Values	N of Above	Ranges
Mean	19.67	25.05	23.55	24.90
Standard Deviation	11.34	9.15	8.89	6.68

TABLE 5

Projected Percentile Rank Equivalents of Selected Scores on 40-Item Tests

Score on 40-Item Test	Percentile Rank of Score			
	Completion	5-Values	N of Above	Ranges
29.223	80	67	74	68
25.617	70	53	59	52
22.543	60	40	45	39
19.674	50	29	33	28
16.918	40	20	23	19
13.731	30	12	13	11
10.125	20	6	7	5

TABLE 6

Reliability Associated with
Various Item Formats

	Completion	5-Values	N of Above	Ranges
Reliability estimates pooled across forms	.572	.465	.432	.423
Reliability adjusted to a 40-item test	.947	.921	.910	.907
Proportion of items required for reliability equivalent to Completion format	1.00	1.54	1.76	1.82