Comparison of Difficulties and Reliabilities of Quantitative Word Problems in Completion and Multiple-Choice Item Formats

Albert C. Oosterhof and Pamela K. Coats
Florida State University

Quantitative word problems were written as parallel completion and multiple-choice items, and were administered to 232 undergraduate students to compare the reliabilities and item difficulties associated with these formats. The multiple-choice options were written using specific numerical responses for each of five alternatives, revised by replacing the fifth option with "none of the above," and also by replacing each of the five responses with ranges of numerical values. Differences in distributions of scores imply a need to reestablish standards if changes are made in the proportions of completion and multiple-choice items included in a test. Findings did not support camouflaging the correct response by using "none of the above" or ranges of numerical values as multiple-choice alternatives. The increased time required to develop and administer a multiple-choice test with reliability equal to that of a completion test suggests use of the latter even in classes with relatively large enrollments.

Authors of educational measurement texts generally favor use of test items that require making a choice among specified alternatives in contrast to items that require the examinee to produce a limited free response. Wesman (1971) recommended against the use of short-answer items, concluding that their superiority over selection-type items is more apparent than real in actual testing situations. Ebel (1979) indicated 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) took a more cautious approach by suggesting that 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 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 included 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.

Background

The literature contains a limited number of investigations comparing completion and various multiple-choice formats in the context of solving quantitative word problems. Wesman and Bennett (1946) used a multiple-choice test battery admin-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 3, Summer 1984, pp. 287–294
© Copyright 1984 Applied Psychological Measurement Inc.
0146-6216/84/030287-08$1.65

istered to nursing school applicants. A portion of subjects was administered a modified form of the test in which the fifth alternative was changed to "none of these." The difficulty and item-test correlation of test items that measured arithmetic skills were, on the average, quite similar for the versions.

Frederiksen and Satter (1953) discussed the development of the U.S. 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) who used items measuring verbal skills, Traub and Fisher administered items rewritten in the multiple-choice format two weeks later. Mean test scores were 3% to 6% lower when mathematics items were written in the multiple-choice format (though the mean scores for multiple-choice items were 60% to 71% higher for the verbal items). 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 were required in the Traub and Fisher study to administer the battery of instruments. Student motivation was recognized as a problem within the experimental conditions.

Other studies have compared multiple-choice and free-response formats in a context other than quantitative word problems. For example, Ward, Frederiksen, and Carlson (1980) contrasted the free-response format with nine-option multiple-choice items for assessing examinees' ability to formulate hypotheses from research data. The two formats were found to have low correlations with each other. Alpha reliability coefficients varied among six scales derived from the test and did not consistently favor either format. Test means could not be compared directly across formats due to the scoring procedures used.

Ward (1982) compared four formats for measuring examinees' ability to identify words representing antonyms, analogies, and missing parts within sentences. Examinees responded by (1) selecting an answer among five alternatives, (2) selecting an answer from a large alphabetized "keylist," (3) providing a single free response, or, (4) providing up to three free responses. The study found no evidence of a unique measurement construct associated with the use of the completion format. Reliabilities were somewhat higher for single-answer completion items than for multiple-choice items used to test antonyms, analogies, and sentence completion skills. Inconsistent findings resulted in similar comparisons between the keylist and multiple-choice completion formats. The keylist format was relatively easy in terms of the percentage of possible points achieved by examinees, whereas the other three formats were roughly compatible with each other but more difficult than the keylist format.

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. To control for retention effects as well as variations in experimental conditions that could occur with separate administrations, alternate test forms containing varying item formats were administered concurrently to groups of examinees equated through random assignment. Multiple-choice foils 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 values were investigated as possible options for reducing the probability of an examinee selecting the correct response without first solving the problem.

Method

An examination in a business finance course was used in the investigation. The examination was developed by the instructor using item construction principles discussed in introductory measurement texts. The test length varied from 34 to 40 items across the academic terms in which the study was conducted. Twelve of these items required the use of quantitative skills to solve problems, and consequently, only this subset of items was used within the present study. Each of the 12 items was written in the following four formats (abbreviated identifiers are given in parentheses):

1. Completion, for which the examinee wrote in the calculated answer.
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. Figure 1 illustrates how an item was adapted to each of the formats.

Four forms of the examination were prepared with the 12 items relevant to the study presented as a consecutive set at or near the end of the test. Table 1 describes how the 12 items appeared in the same order within each form, but appeared in different formats across the four forms. For example, the first of these 12 items appeared as the following: A completion item in Form A, a multiple-choice item with five numerical values for options in Form B, an identical item in Form C except that Option E was changed to ''None of the above,'' and a multiple-choice item in Form D with ranges instead of point values for each of the five options. Table 1 indicates how triads of items used an A, C, or E as the correct multiple-choice alternative, but not necessarily in that order. Consequently, the distribution of keyed responses was held constant across forms of the test, and ''none of the above'' (Option E) was used as a correct response once within each triad of items.

The 12 items were administered to 232 undergraduate business majors as part of a course examination in each of three academic terms. Students were encouraged to respond to all items, and test scores represented the number of correct responses. The four forms of the test 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.

Figure 1
Illustration of an Item Adapted
to the Four Formats

Item Stem

If Internal Rate of Return equals 11%, 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

Completion: ANSWER ____________

5-Values:
A. $268.13
B. $294.00
C. $313.07
D. $326.00
E. $358.00

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

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

Table 1
Format of Items and Numbers of Examinees Administered Each Form

Item	Key	Form A	Form B	Form C	Form 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				
Examinees		60	59	57	56

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 correct. The answer forms were then machine scored with all items scored dichotomously.

Item difficulties (p values) were calculated separately for the 12 items written in each of the four formats. The mean difficulty was then established for the group of 12 items written in each 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. The KR-20 reliability coefficient was calculated for each triad of items within each of the four item formats, and the average reliability was derived for each format.

Results

Item difficulties for the 12 items within each of the four formats are listed in Table 2. (To facilitate analysis, values associated with a given format are grouped into a single column, though triads of items written in a given format were distributed across four forms administered to groups equated through random assignment.) The items incorporated in the investigation were mostly of moderate difficulty with the middle 50% of the values ranging between .48 and .70. Even with a somewhat restricted range of difficulties, correlations between rankings of item difficulties within the four item formats ranged from .72 to .91. No significant differences were observed in level of performance across the three terms during which data were collected. For example, the mean scores on the 12 items included in Form A were 6.0, 6.6, and 7.2 when administered to 21, 12, and 27 students, respectively. Fewer than 3% of the students omitted a response to an item written in the multiple-choice format. Fewer than 10% omitted a response to a completion item.

Completion items were consistently the most difficult, with the three multiple-choice formats being of near equal difficulty. The Friedman (1937) index indicates that the consistency of rankings of item difficulties across the four item formats observed for each item in the present study would occur with a probability < .001 if item difficulty was independent of item format. Subsequent paired comparisons using Wilcoxon's signed rank test resulted in probabilities < .02 between the completion and each of the multiple-choice formats, and probabilities > .05 for comparisons among the multiple-choice formats.

Table 3 illustrates how substituting ''none of the above'' as an option generally made the item more difficult than other multiple-choice formats, almost all of this increased difficulty occurring when ''none of the above'' was the correct answer.

The average reliability associated with each of the four item formats is presented in Table 4. KR-20 estimates of reliability based on triads of items and then averaged 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 a slightly higher reliability than N of Above and Ranges.

Discussion

Differences in item difficulty were 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 completion format is used. This is inconsistent with the findings of Traub and Fisher (1977). This may have resulted from subjects in the Traub and Fisher study being administered the multiple-choice items as a retest using the same item stems used in the completion items, although the two administrations were separated by a two-week interval. Another possible cause may have been motivational differences. Subjects in the present study were administered the items as part of a significant class examination.

It is probable that examinees will rework a problem presented in the 5-Values format if the worked solution is inconsistent with all five alternatives. If

Table 2
Item Difficulties for 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	.646	.717
6	.695	.789	.667	.750
7	.404	.500	.317	.475
8	.439	.542	.633	.695
9	.702	.875	.883	.831
10	.542	.550	.678	.544
11	.562	.600	.729	.667
12	.188	.300	.119	.368
Mean	.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

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

a solution consistent with an alternative cannot be obtained, the examinee will likely choose the alternative perceived most consistent with the obtained solution to the problem. (Even if the student is unable to consciously eliminate any of the alternatives, the option perceived to be most plausible rather than one chosen by a random guess is likely to be the selected response when a penalty is not imposed for giving wrong answers.) Only if the foils are able to encompass a high proportion of incorrect solutions and the correct solution is perceptually deviant from probable incorrect solutions would a 5-Values 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 the present examinees were wary of using this alternative unless they were confident of their calculated solution.

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 it may permit selection of the keyed response by obtaining a nearly correct solution for the wrong reason. Ranges will probably also 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. If, for example, tests of 40 quantitative world problems were constructed and average item means, variances, and covariances were consistent with those observed in the present study for the respective item formats, then resulting means and standard deviations of the 40-item tests could be projected. Assuming normal distributions of scores for each of the tests, percentile equivalents across the four formats can be estimated. For example, a projected score of 19.7 would represent the 50th percentile for Completion items, but only 29%, 33%, and 28% of the examinees would be expected to score below this score when administered corresponding tests using the respective multiple-choice formats. Distributions of scores may not be normal as assumed here, and the relative difficulties of respective item formats are likely to vary as the

Table 4
Reliability for Each Format

	Completion	5-Values	N of Above	Ranges
Reliability estimates averaged across triads	.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

nature of examinees and content measured change. However, differences in means and variability of test scores resulting from varying item formats probably are sufficiently significant to merit reestablishing standards if meaningful changes are made in the portions of completion and multiple-choice items included in tests involving quantitative word problems. The resulting differences in distributions of test scores similarly suggest that mastery standards for ''criterion-referenced'' tests should include specification of the type of item format to be used.

The reliability of all four item formats is respectable. Table 4 indicates that if the average reliability obtained from triads of items was adjusted with the Spearman-Brown formula to represent a more typical test length of 40 items, then all formats would result in high reliabilities. However, as also indicated in Table 4, a significant proportion of additional multiple-choice items would be required to obtain reliability equivalent to the Completion format. Assuming item correlations between items equal to those observed for the respective item formats in the present study, 62, 70, and 73 items of the respective multiple-choice formats would be required to match the reliability of 40 Completion items. Had the study involved items or students that resulted in lower reliability estimates, a smaller proportion of multiple-choice items would be required to match the reliability of the completion items. For example, if average inter-item correlations among completion items was .10 and the ratio of average inter-item correlations of Completion and Ranges remained constant, then the ratio of Completion to Ranges items required for equivalent reliability would become approximately 1 to 1.5. An 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.

Minimal advantage was found, 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 completion format when efficiency of scoring is not an overriding concern, such as when the test is to be administered to large numbers of examinees. Generalization from this research context to other measurement settings must be performed cautiously since relative difficulty and reliability of multiple-choice items are dependent on the choice of distractors.

References

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

Frederiksen, N., & Satter, G. A. (1953). The construction and validation of an arithmetic computation test.

Educational and Psychological Measurement, 13, 209–227.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32,* 675–701.

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

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

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

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

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

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6,* 1–11.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement, 17,* 11–29.

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

Wesman, A. G., & Bennett, G. K. (1946). The use of ''none of these'' as an option in test comparison. *Journal of Educational Psychology, 37,* 541–554.

Author's Address

Send requests for reprints or further information to Albert C. Oosterhof, 307 Stone Building, Florida State University, Tallahassee FL 32306, U.S.A.