DOCUMENT RESUME

ED 064 361 TM 001 529

AUTHOR Oosterhof, Albert C.; Glasnapp, Douglas R.
TITLE Comparative Reliabilities of the Multiple Choice and True-False Formats.
PUB DATE Apr 72
NOTE 5p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Comparative Analysis; *Guessing (Tests); *Multiple Choice Tests; Ratios (Mathematics); *Student Evaluation; *Test Construction; Test Reliability
IDENTIFIERS *True False Tests

ABSTRACT

The present study was concerned with several currently unanswered questions, two of which are: what is an empirically determined ratio of multiple choice to equivalent true-false items which can be answered in a given amount of time?; and for achievement test items administered within a classroom situation, which of the two formats under consideration result in greater reliability per unit of testing time? Subjects were 101 undergraduates enrolled in one section of an introductory measurements course. Forty multiple choice items were selected on the basis of their relationship to stated course objectives and according to their ability to discriminate between levels of achievement. Data from this research indicate that true-false items, particularly those items which are in fact true, result in a less reliable test than had a four-option multiple choice format been used. It also appears that when the correction for guessing formula is applied in order to equalize scores relative to items correctly answered on a pure chance basis, the multiple choice item is the easier of the two formats to answer, with items keyed true easier than those keyed false with regard to the true-false format. (Author/LS)

Comparative Reliabilities of the Multiple Choice and True-False Formats[1]

Albert C. Oosterhof and Douglas R. Glasnapp

The University of Kansas

Considerable discussion has taken place among measurement specialists regarding the virtues of multiple choice versus true-false test item formats. Recent contrasting examples might include "...the advantages attributed to (true-false items) are not, unfortunately, very valid.... (Gronlund, 1971, p. 160)", and "...a few (test specialists) see special virtues of efficiency and ease of preparation in (true-false items) and advocate their wide use (Ebel, 1971, p. 1)." The most obvious limitation of true-false relative to multiple choice test items is the degree to which the former is subject to guessing. Several studies have shown that the reliability of a test is directly related to the number of choices per item (Remmers, Karslake, and Gage, 1940; Lord, 1944; Carroll, 1945; Plumlee, 1952). Similarly, it would be expected that a multiple choice test would have greater reliability than a true-false test if the number of items were held constant. However, since a greater number of true-false items can be administered per unit time, it is possible that in a given amount of time, the increased number of true-false items administered would allow for greater reliability and more efficient sampling of content objectives than had a multiple choice format been used.

Using 88 multiple choice items from a published test in natural science, Ebel (1971) compared formats by rewriting each multiple choice item as a parallel true-false item. Two forms, each consisting of 44 multiple choice and 44 true-false items, were developed. Reliabilities (K.R. 20) were computed for the multiple choice and true-false sections of both forms, and assuming that two true-false items could be answered per multiple choice item, the Spearman-Brown formula was used to predict the reliability of an 88 item true-false test. For the first form, this adjusted reliability was greater than the reliability obtained for the multiple choice section of the test, however the inverse was true with respect to the second form.

The present study was concerned with several currently unanswered questions. First, what is an empirically determined ratio of multiple choice to equivalent true-false items which can be answered in a given amount of time? Second, for achievement test items administered within a classroom situation, which of the two formats under consideration result in greater reliability per unit of testing time? Third, what is the relative reliability of true true-false and false true-false items when compared to multiple choice items? Fourth, what ratio of multiple choice to equivalent true-false items is necessary for producing equal reliability coefficients? Lastly, after equating for differences in the effect of guessing, what is the relative difficulty of the different formats?

---

[1]A paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 3-7, 1972.

Method

One-hundred one undergraduates enrolled in one section of an introductory measurements course served as subjects (Ss). Forty multiple choice items were selected from an item pool on the basis of their relationship to stated course objectives, and according to their ability to discriminate between levels of achievement. Only items which consisted of one correct option and three independent and incorrect options were used. Each multiple choice (MC) item was rewritten as a true-false item keyed true (Tf) by combining the stem and correct option, and also as a true-false item keyed false (tF) by combining the stem and the best discriminating incorrect option. An example of a MC item and corresponding Tf and tF items is provided in Illustration 1. The total of 120 items were used as the final course examination for all Ss. Part 1 of this exam consisted of the 40 MC items whereas Part 2 contained the 80 Tf and tF items. Each pair of true-false questions that were generated from the same MC item were randomly assigned to the first or second set of 40 items to Part 2. The position of each true-false item was then randomly assigned within each of these two sets. Fifty-one of the Ss began with Part 1 of the exam while the remaining Ss began with Part 2, both groups completing all 120 items. At the end of 40, 80, and 120 items, the Ss were requested to record the number of minutes required to reach these respective points in the exam, the elapsed time being indicated on the front board.

Separate reliabilities were computed from the 40 MC, Tf, tF and mixed true-false (Mtf) items. The reliability of all 80 Mtf items (Tf + tF items) was obtained and using the Spearman-Brown formula, the reliability of a 40 Mtf item test was calculated in order to keep test lengths equal for comparative purposes. Average elapsed times were computed for MC and true-false items (times for Tf and tF items could not be computed separately since these items were intermixed, and for purposes of this study their times were assumed to be equal). Using the Spearman-Brown formula the reliabilities of the Tf, tF, and Mtf items were adjusted for differences in time required to answer MC items. There reliabilities were also adjusted using the 2:1 ratio incorporated by Ebel (1971). Again using the Spearman-Brown formula, the required ratio of Tf, fT, and Mtf to MC items required for equivalent reliabilities was computed. Modifying the respective Ss scores with the correction for guessing formula, a repeated measures ANOVA design was used to compare the difficulties of MC, Tf and tF items.

Illustration 1. Sample Items

A major advantage of individual intelligence tests over group tests is that

A. the standardization group is usually larger
*B. information other than the test score can be obtained
C. the method of scoring is more objective
D. they must be administered by skilled examiners

T F Individual intelligence tests are superior to group intelligence tests in the sense that individual tests provide more information.

T F Relative to scoring procedures, individual intelligence tests are superior to group intelligence tests in that individual tests are more objectively scored.

Results

Table 1 provides descriptive statistics related to sections of the exams composed of MC, Tf, and tF items. These indices are also given for the combined true-false (Mtf) items, and for the test as a whole. Discriminations are point biserial correlations between Ss scores on individual items and total test scores. Reliability coefficients were determined using the Kuder-Richardson formula No. 20. With the exception of the reliability coefficients, the information contained in this table is for background information only.

The average amount of time required to answer MC items was 1.18 minutes, while the average time was .68 minutes for true-false items. This resulted in a ratio of 1:1.73 multiple choice items to true-false items that were answered per unit time. Table 2 provides the reliabilities before adjustment associated with items of each format, and corresponding reliabilities after adjustments using the Spearman-Brown formula. Reliability associated with Mtf items was adjusted from 80 to 40 items for comparative purposes. Each true-false format was adjusted, on account of different amounts of time required to answer multiple choice and true-false items, to represent tests 1.73 times the length of 40 items, and similarly to tests twice as long as 40 items. Table 2 also indicates the number of test items of each item format which would have been required per multiple choice item in order to establish equivalent reliabilities.

The average adjusted scores obtained with the MC, Tf, and tF items were 19.91, 14.87 and 12.34 respectively. The hypothesis of equal means was rejected ($F=45.99$; $df=2,200$; $p<.01$). Post hoc procedures utilizing the Scheffe technique demonstrated that each mean was significantly different from the other two ($p<.01$).

Table 1

Data on Various Item Formats

Item Format	MC	Tf	tF	Mtf	All Items
Number of Items	40	40	40	80	120
Mean No. Correct	24.93	27.44	26.17	53.60	78.53
Standard Deviation	6.18	3.80	4.66	7.07	12.38
Median Difficulty	.640	.695	.640	.645	.645
Median Discrimination	.350	.195	.245	.230	.265
Reliability	.816	.503	.648	.702	.856

Table 2

Comparison of Reliability Coefficients

Item Format	MC	Tf	tF	Mtf
Reliabilities:				
Unadjusted	.816	.503	.648	.702
Adjusted to 40 items	.816	.503	.648	.541
Adjusted for time ratio of 1:1.73	.816	.636	.761	.671
Adjusted for time ratio of 1:2	.816	.669	.786	.702
Number of items per MC item required for equivalent reliability	1.00	4.38	2.41	3.75

Discussion

Data from this research have indicated that true-false items, particularly those items which are in fact true, result in a less reliable test than had a four-option multiple choice format been used. This relationship held true even when differences in time needed to answer the respective formats were taken into account. The data suggested that approximately two and one-half to four and one-half as many true-false as multiple choice items were necessary in order to produce equivalent reliabilities, this ratio being greater than the frequency with which true-false items would be answered relative to multiple choice items. This would have been the situation even had the ratio of true-false to multiple choice items answered per unit time been 2:1. This supports the conclusion that if the true-false format were used in lieu of multiple choice items for achievement tests administered within a classroom situation, the increase in content sampling would be accomplished at the sacrifice of reliability.

However, one might infer that since several of the items written in the true-false format and used in the present study obtained discriminations (point-biserial correlations) within the .45 to .55 range, that with time, it would be possible to develop a test consisting entirely of highly discriminating true-false items, whose resulting reliability would consistently rival a parallel test using the multiple choice format. But it does appear that such a possibility lies closer to the domain of standardized tests where extensive item revision is more common than with the development of teacher-oriented instruments.

It also appears that when the correction for guessing formula is applied in order to equalize scores relative to items correctly answered on a pure chance basis, the multiple choice item is the easier of the two formats to answer, with items keyed true easier than those keyed false with regard to the true-false format. Implications of these results when using multiple choice as opposed to true-false items, or vice versa, for formative or summative evaluation in a mastery learning model are evident. Depending on the type of item format used, the number of objectives indicated as mastered would differ.

References

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 1945, 10,1-19.

Ebel, Robert L. Comparative effectiveness of true-false and multiple choice achievement test items. Paper presented at the 55th Annual Meeting of the American Educational Research Association, February, 1971.

Gronlund, Norman E. Measurement and evaluation in teaching. (2nd ed.) Macmillan, 1971.

Lord, F. M. Reliability of multiple choice tests as a function of number of choices per item. Journal of Educational Psychology, 1944, 35, 175-80.

Plumlee, L. B. The effect of difficulty and chance success on item-test correlation and on test reliability. Psychometrika, 1952, 17, 69-86.

Remmers, H. H., Karslake, R., and Gage, N. L. Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown Prophecy Formula. Journal of Educational Psychology, 1940, 31, 583-90.