DOCUMENT RESUME

ED 185 097 TM 800 101

AUTHOR Weiten, Wayne
TITLE Relative Effectiveness of Single and Double Multiple-Choice Questions in Educational Measurement.
PUB DATE [Sep 79]
NOTE 13p.; Paper presented at the Annual Meeting of the American Psychological Association (87th, New York, NY, September 1-5, 1979).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; Higher Education; *Item Analysis; *Multiple Choice Tests; Test Construction; *Test Items; Test Reliability; Test Results; Test Validity
IDENTIFIERS Item Discrimination (Tests)

ABSTRACT

Two different formats for multiple-choice test items were compared in an experimental test given in a college class in introductory psychology. In one format, a question or incomplete statement was followed by four answers or completions, only one of which was correct. In the other format, the double multiple-choice version, the same questions were used together with many of the same answers: but any number of these answers might be correct and the second set of choices required the test taker to choose a response that showed which answer or combination of answers was right. The results of this experiment showed the double multiple-choice format to be more difficult and less discriminating than the standard format. The two 20-item subtests made up of double multiple-choice items were significantly less reliable than the parallel 20-item subtests made up of standard format items: however, there were no differences in the validities of the two formats in predicting the final course grades. (CTM)

RELATIVE EFFECTIVENESS OF SINGLE AND DOUBLE MULTIPLE-CHOICE QUESTIONS IN EDUCATIONAL MEASUREMENT[1]

WAYNE WEITEN

COLLEGE OF DuPAGE AND UNIVERSITY OF ILLINOIS AT CHICAGO CIRCLE

RUNNING HEAD: DOUBLE MULTIPLE-CHOICE QUESTIONS

[1]The author is indebted to Nate Clark for making subjects available and to Bruce Korth for his advice regarding this study. Portions of this article were presented at the 1979 meeting of the American Psychological Association in New York.

ABSTRACT

Double multiple-choice questions include two sets of options; from the latter set the subject must select the option which identifies the correct or best collection of options in the first set. A comparison of double as opposed to single multiple-choice questions was made on a classroom psychology exam taken by 47 undergraduates. Forty pairs of matched single and double format questions were prepared and 20 items of each type appeared on each of two forms of the exam. Significant differences were observed between the two question types in regard to item difficulty, item discrimination, and internal reliability but not concurrent validity. The results were interpreted as suggestive that double multiple-choice questions may increase variance attributable to certain subject characteristics other than content mastery.

RELATIVE EFFECTIVENESS OF SINGLE AND DOUBLE MULTIPLE-CHOICE QUESTIONS IN EDUCATIONAL MEASUREMENT

The present study examined the effects of employing "double multiple-choice" questions as opposed to conventional "single multiple-choice" questions in educational measurement. The term "double multiple-choice" is used to describe questions of the form seen in the example below which was drawn from the 1977-78 Law School Admission Bulletin and LSAT Study Guide.

14. Which of the following can be inferred from the graphs?
    - I. The Jewish population of North America is larger than the Jewish population of any other continent.
    - II. Of the six continents, Asia has the greatest number of people who profess no religion.
    - III. South America has a larger population of Roman Catholics than any other continent.

    A. None
    B. I only
    C. III only
    D. I and III only
    E. I, II, and III

As can be seen above, double multiple-choice questions include two sets of options. From the latter set, the subject must select the option which identifies the correct or best collection of options listed in the first set. In contrast, a single multiple-choice question involves only one set of options.

It is apparent that double multiple-choice questions share some continuity with the rather common practice of using options such as "all of the above" in conventional multiple-choice questions with only one set of options. However, the practice of including two complete sets of options appears to represent a substantial departure from conventional item writing

procedures which mertis empirical examination.

Although the author has no systematic evidence, the popularity of double multiple-choice questions appears to be increasing and questions of this type now appear on important admissions tests (e.g. LSAT) and licensing examinations (e.g. Illinois State Board Test Pool Exam for Registered Nurses). An extensive search of the educational measurement literature uncovered no evidence on the relative merits of double as opposed to single multiple-choice items. This dearth of empirical evidence is not surprising. Although there is a substantial literature on rules for writing multiple-choice questions (cf. Gronlund, 1965; Wood, 1960) this literature has virtually no empirical data base. Thorndike and Hagen (1955) provided an erudite description of the situation almost 25 years ago which unforunately is still accurate today.

> The point we wish to make is that we do not have a science of test construction. The guides and maxims we will offer (for item construction) are not tested out by controlled scientific experimentation. Rather, they represent a distillation of practical experience and professional judgment. (p. 50).

The reasons for this paucity of research on item construction are unclear. The problem of writing sound test items obviously has considerable practical relevance to important endeavors in educational measurement, and does not seem to be impeded by any particularly formidable medthodological roadblocks.

While there apparently is no empirical evidence, there are a number of admittedly speculative reasons for suspecting double and single multiple-choice items might be differentially effective. First, certain characteristics of the double format questions are inconsistent with some of the conventional item writing guidelines (Ebel, 1965; Gronlund, 1965). For instance, double format questions tend to conflict with conventional wisdom which puts a premium on item brevity and cautions against writing multiple-choice questions

in which the options function as a series of true-false propositions. Second, one could argue that the lengthier and more complex double format questions might be more sensitive to individual differences in reading ability or reasoning skill. Since an educational test is usually designed to measure some content knowledge, additional score variation attributable to differences in reading or reasoning skill would represent an increment in error variance. Third, there is evidence that subjects' exam performance may be influenced by their test-wiseness (Millman, Bishop, & Ebel, 1965; Rowley, 1974). It seems reasonable to conjecture that the more elaborate double format items may be more sensitive to these individual differences in test-taking skill.

Cognizant of the possibilities outlined above, the present study compared single and double multiple-choice questions in regard to difficulty, item discrimination, reliability, and validity. Based on anecdotal evidence gleaned from students' common complaints that double format items are excessively difficult, it was hypothesized that item difficulty indices would be lower for the double format questions than for the single format questions. Based on reasoning that double format items may introduce additional sources of error variance, it was hypothesized that the double multiple-choice questions would yield lower item discrimination indices than the conventional questions and that subtests consisting of these double format items would diplay lower reliability and validity than the matching single format subtests.

METHOD

Subjects and Procedure

Subjects were undergraduates enrolled in a large introductory psychology course who were required to participate in an experiment of their choice. Forty-seven students (28 female, 19 male) elected to participate in the present study. One week after the first of two exams in the course, the subjects were

assembled outside of class and were administered one of two experimental exams covering the same reading assignment as the first exam in the course. Thus, subjects were tested on actual course material that they had recently studied. Subjects were unaware of the nature of the study until they arrived for the experimental session. At that time, the nature of the experimental test was explained, and they were informed that the results would not have any effect upon their course grade. The two forms of the experimental test were passed out alternately so that 24 subjects responded to Form A and 23 subjects responed to Form B.

Test Construction

The items for the two forms of the experimental test were drawn from the instructor's manual designed to accompany the required course textbook, Psychology Today: An Introduction (1975). The selection of items from the pool of multiple-choice questions available for the assigned chapters was not random. All items previously used by the course instructor (not the author) on the actual mid-term exam had to be eliminated. Furthermore, since all selected items were to be rewritten into a double multiple-choice format, the author had to exercise some subjective judgment in choosing items which could be sensibly transformed into the double format. For each of the conventional questions drawn from the test item pool, a similar double multiple-choice item was constructed. Matched items in the two formats had identical stems, and in most cases the single set of options in the conventional question was exactly the same as the first set of options in the double multiple-choice version. Exact identity for all matched questions was impossible to achieve because the conventional questions each had only one correct option, whereas at least some of the double format questions had to have more than one correct option in the first set so that the correct answer in the second set could involve a collection of options.

Two examples of matched items can be seen below. In the first example, there is exact identity between the only set of options in the single format version and the first set of options in the double format version. In the second example, there is a slight difference between the first set of options in the double format and the only set of options in the single format.

Example One

Single multiple-choice version

Which of the following approaches postulates that children are taught sex roles through conditioning and modeling?

a. Psychoanalytic
*b. Social-learning
c. Cognitive
d. Humanistic

Double multiple-choice version

Which of the following approaches postulates that children are taught sex roles through conditioning and modeling?

1. Psychoanalytic
2. Social-learning
3. Cognitive
4. Humanistic

a. 1 only
*b. 2 only
c. 2 and 3 only
d. 2, 3, and 4 only

Example Two

Single multiple-choice version

The study of psychology is important to everyone because it provides:

a. a new perspective from which to view daily events of life
b. insight into one's own behavior
c. practical information
*d. all of the above

Double multiple-choice version

The study of psychology is important to everyone because it provides:

1. a new perspective from which to view daily events of life
2. insight into one's own behavior
3. practical information
4. solutions to everyone's problems

a. 1 only
b. 1 and 3 only
*c. 1, 2, and 3 only
d. 1, 2, 3, and 4

Care was exercised in the construction of questions with discrepancies such as that illustrated in the second example so as to maintain the essential character of the original question. Thus, the only significant disparity between matched items was the addition of the second set of options in the double multiple-choice format.

Two sets of 40 matched items were developed in this manner. Twenty items were then randomly assigned to appear in single format on Form A and in double format on Form B. The remaining twenty items were presented in single format on Form B and in double format on Form A. The order of items was identical for the two forms of the experimental test and corresponded to the order of items in the instructor's manual (as well as the order of coverage in the text).

Dependent Variables

Four dependent variables were examined. Item difficulty indices represented the proportion of subjects correctly answering an item, so that lower figures were indicative of greater difficulty. Since both types of questions were ostensibly legitimate measures of content mastery, item-whole point-biserial correlations were computed to provide item discrimination indices. To assess internal reliability, each 40 item test was divided into a pair of 20 item subtests composed exclusively of one type of item format. This necessitiated two comparisons, one between the single items on Form A and the matched double

items on Form B, and one between the double items on Form A and the corresponding single items on Form B. Concurrent validity was measured by correlating each subject's single and double subtest scores with that subject's cumulative point total for the entire course.

RESULTS AND DISCUSSION

Item Difficulty

Mean item difficulty for the 40 single multiple-choice questions was .633 while the 40 double multiple-choice questions yielded a mean of .575. A directional t test for differences between correlated means revealed that this difference was significant ($t = 1.74$, $df = 39$, $p < .05$). The observed difference is consistent with the common complaint of examinees that the double format items are more challenging than traditional single format items. This finding also makes sense in view of the seemingly more complex nature of double multiple-choice questions.

However, it should be pointed out that disparity in item difficulty is not crucial to the issue of differential effectiveness in educational measurement. Although double format questions may tend to be more difficult than comparable single format questions, it should be stressed that item difficulty can be manipulated (through the judicious selection and careful writing of options) to a desired level or optimal range within either format. Nonetheless, the data suggest that students' dislike of double format questions is not simply a matter of resisting the unfamiliar. These questions do appear to confront the examinee with a more difficult task.

Item Discrimination

The point-biserial correlations used to estimate item discrimination do not represent interval data and they usually are characterized by a decidedly skewed distribution. Therefore, a non-parametric test (Wilcoxon's T) was used to compare the distributions of item discrimination indices. Median

discrimination for the single and double items was .34 and .23 respectively. The Wilcoxon signed-ranks test indicated that the single-format items discriminated significantly better ($T = 254.5$, $p < .05$) than did the double-format items. Insofar as the experimental tests measured course knowledge, the disparity in item discrimination suggests that the single-format questions did a better job than the double-format questions in distinguishing between well informed and poorly informed students.

Reliability

The significance of the two comparisons of the 20 item subtests in regard to reliability was assessed with Feldt's (1969) W which is approximately distributed as F. In the first comparison, KR-20 reliability for the A-single subtest was .71 as compared to .38 for the B-double subtest ($W = 2.12$, $df = 22, 23$; $p < .05$). In the second comparison, KR-20 reliability was .58 for the B-single subtest and .45 for the A-double subtest ($W = 1.32$, $df = 23, 22$, $p > .05$). Thus in both cases, internal relilability was noticeably lower for the double-format subtest, although only one of the differences was significant.

These estimates of internal reliability are largely a function of test length and the homogeneity of test content. Since the matched subtests were of the same length and involved the same sampling of the content domain, the lower internal reliability observed on the double format subtests suggests that the double multiple-choice questions may have increased the variance attributable to factors which are independent of subjects' content mastery. The low internal reliability displayed by the double format subtests, coupled with the observed difference in item discrimination, seems consistent with the conjectural analysis that the double format questions may produce greater variance due to subject differences in reading and reasoning skills or test-wiseness.

Validity

No significant difference was observed between the two formats in regard

to concurrent validity; both the single-format and the double-format subtests correlated .54 with students' actual course totals. This lack of a difference in concurrent validity seems inconsistent with the differences observed in regard to item discrimination and reliability. However, this apparent inconsistency may have a reasonable explanation. The double-format questions may introduce additional sources of variance which happen to be highly related to the criterion variable of course knowledge. For instance, if reading and reasoning skills are highly correlated with course mastery (an intuitively plausible assumption), then additional variance attributable to these subject characteristics would not necessarily lower the validity of the double-format subtests.

Summary

Overall, the pattern of results suggests that there may be some interesting differences between double and single-format questions which merit futher research. The implicit assuption that the two formats are equivalent was not supported by the data. In addition to the expected, and relatively innocuous, discrepancy in difficulty level, more distrubing differences in item discrimination and test reliability were found. These differences suggest that double multiple-choice questions may generate additional variance in test scores which is not attributable to differences in content mastery. Insofar as this additional variance may largely reflect differences among subjects in important cognitive skills such as reading and reasoning, it would not be particularly problematical on aptitude tests such as the LSAT. In contrast, on classroom tests or licensing exams, which are intended to measure mastery of specific information, these sources of variance would clearly represent an increment in error variance. However, in view of the failure to observe a difference in concurrent validity on the classroom test used in the present study, any alarming assertions about the differential effectiveness of the two formats would be premature.

REFERENCES

Ebel, R.L. Measuring Educational Achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965.

Feldt, L.S. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 1969, 34, 363 - 373.

Gronlund, N.E. Measurement and Evaluation in Teaching. New York: Macmillan, 1965.

Millman, J., Bishop, C.H. & Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 1965, 25, 707 - 726.

Psychology Today: An Introduction (3rd Ed.). New York: Random House, 1975.

Rowley, G.L. Which examinees are most favoured by the use of multiple choice tests? Journal of Educational Measurement, 1974, 11, 15 - 23.

Thorndike, R.L. & Hagen, E. Measurement and Evaluation in Psychology and Education. New York: John Wiley, 1955.

Wood, D.A. Test Construction. Columbus, Ohio: Charles E. Merrill, 1960.