DOCUMENT RESUME

ED 110 498 TM 004 771

AUTHOR Diamond, James J.; Williams, David V.
TITLE Demand Characteristics of Multiple-Choice Items.
SPONS AGENCY Spencer Foundation, Chicago, Ill.
PUB DATE [Apr 75]
NOTE 5p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D.C., March 31-April 2, 1975); For a related document, see TM 004 765

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Classification; *Cognitive Processes; Criterion Referenced Tests; Graduate Students; Memory; *Multiple Choice Tests; *Recall (Psychological); *Response Style (Tests); Test Construction; Testing

ABSTRACT

Thirteen graduate students were asked to indicate for each of 24 multiple-choice items whether the item tested "recall of specific information," a "higher order skill," or "don't know." The students were also asked to state their general basis for judging the items. The 24 items had been previously classified according to Bloom's cognitive-skills hierarchy. The results of the study supported the hypothesis that the examinees' judgment of the cognitive process being measured by each item is influenced by the structure of the item--for example, stem length. (Author)

Demand Characteristics of Multiple-Choice Items*

James J. Diamond
University of Pennsylvania

and

David V. Williams
Ithaca College

This study is an outgrowth of a study of memory vs. inference processes (as seen in the report of multiple-choice item-solution processes). It examines the hypotheses with which a student may approach a test item, whether or not (s)he may know the answer.

Observation of the classroom testing situation suggests that students often have hypotheses about the processes being elicited by a multiple-choice item (e.g. whether it measures "recall" or is a "thought" question), which presumably enters into the student's strategy for either answering or guessing. Theoretical considerations, too, suggest that since an achievement test can be viewed as an experiment (with instruction as the independent variable, and the score as the dependent variable) the concept of "demand characteristics" might be applied to the situation. (This term is applied from field of social psychology where it refers to characteristics of an experiment other than the treatment which affects the subjects responses).

In the case of multiple-choice items, students are known to try to "psych-out the professor." If a student doesn't know the answer to an item (s)he can attempt to deduce it using strategies that fall under the rubric of "test-wiseness." In so doing the student may well hypothesize whether the item is testing recall of specific material or some "higher order" cognitive skill. Despite current emphasis on so-called criterion-referenced tests, the criterion behavior is typically never observed. Rather, approximations of that behavior are elicited through a number of techniques including multiple-choice tests. An understanding of the role of hypotheses about the intended process of an item may aid in the elimination of disfunctional "demands" thus presenting a means of increasing the criterion validity of such test scores.

The purpose of the present study was to determine whether students can reliably categorize items as intending to test recall or higher order skills based simply on the structure of the items. In addition information regarding the cues employed for this classification were sought.

Method

Students from Ithaca College and the University of Pennsylvania were asked to read two sets of twenty-four multiple-choice items; one set from a module dealing with glaciers, the other concerning the periodic table of atomic numbers (Diamond & Williams, 1972). The students read only the items and not the associated reading passages. Each was asked to indicate whether, in his/her judgment, the item required the recall of specific information or some higher-order skill. The students were then asked to write down their basis for judging two types of items.

---

* Paper presented at meetings of the National Council for Measurement in Education, 1975. This research was supported by a grant from the Spencer Foundation.

The sets of 24 items had been originally written and judged to reflect Bloom's (1956) cognitive-skills hierarchy of Knowledge, Comprehension, Application, and Analysis (by Kropp et al. 1956). Six items of each type were randomly arranged in booklet form, one item per page.

A correct classification of an item was defined as at least majority agreement with the judges' previous classification. ("Recall" corresponded with Knowledge and "higher-order skill" with any of the other three categories.) Internal consistency measures of the ratings were calculated by assigning a "score" of 1.0 if the student's item classification was the same as that of the judges, and 0.0 otherwise. "Items" was treated as a repeated-factor and the mean-square for subjects x items as an estimate of error variance.

Results

The data for glaciers and atomic structure were very similar, and so are considered jointly.

Number of Correct Classifications/Item

	Glaciers Items		Atomic Structure Items	
	Recall	H.O.S.	Recall	H.O.S.
Penn	5	11	6	12
Ithaca	5	11	6	10
Internal Consistency	.52	.79	.65	.75

Students were clearly able to correctly classify the Recall items. The lowest proportion of correct classification was 16/26 for one of the glacier items; all other proportions were higher.

The written comments indicate that the structure of the item was the basis for their judgments.

For example:

> " "It usually depended on the wording of the question. Certain key words helped me to decide--'relationship,' 'approximately,' 'plausible reason,' 'assumption,' 'best estimate,' 'might be'--which were HOS questions. The R questions seemed to be more direct with direct answers."

"R items seemed to be the ones you could answer immediately from having memorized them. HOS items you usually knew a few facts from memorization but were asked to use all this knowledge and come up with the best answer. More time is needed to answer this type."

"'Can best be described' questions are HOS because the student must organize the information from the passage into a logical answer. Best answer = best prediction = HOS."

Discussion

Accurate classification of Knowledge, Application, and Analysis items was obtained. In fact, only 7 of the 36 items in these areas were incorrectly dichotomized; whereas 9 of the 12 Comprehension items were so rated. This finding adds credence to the structure hypothesis in that the comprehension items that were not of the "can best be described" variety are structurally indistinguishable from recall items. For example, consider the following item:

The serious study of glaciers began about the time of the:

A Civil War.
B Golden Age of Greece.
C French Revolution.
D discovery of America.

The passage on which this item is based says that "Glaciers have been studied seriously for a little more than 100 years." Since the students in this study did not read the passages, the item above can clearly appear to be testing recall of specific information. As indicated, however, there was one item on glaciers presumably testing recall of specific information in which part of the stem used the word "best." This item was presumed by the students to be testing some higher-order skill when in fact the test constructor intended it to measure recall only. Thus, inducing an appropriate "process demand" or "set" may be an important, but neglected, aspect of item construction.

The results of this preliminary study should be viewed in the context of the "extra cognitive" aspects of multiple-choice testing, that is, the situational and structural factors--affecting the percent passing an item and yet not clearly related to the processes the item is intended to measure.

The results of this study seem important in criterion-referenced measurement. If the purpose of such a test is to yield a score which can indicate the behaviors in which students have engaged, then, the process expectancy that the item's structure creates may be a source of systematic measurement error affecting the validity of inferences made from these scores. Systematic understanding of the relation between choice behavior influenced by expectancy (via structure) on one hand and cognitive processes (via content) on the other appears essential for the construction of adequate criterion-referenced items capable of distingusihing between inference and recall.

General References

Bloom, B.S. (ed.) Taxonomy of educational objectives, Handbook I: cognitive domain. New York: David McKay & Co., 1956

Bloom, B.S. & Broder, L.J. Problem-solving processes of college students: an exploratory investigation. Supplementary Educational Monographs, 1950, No. 73.

Campbell, A.C. Some determinants of the difficulty of non-verbal classification items. Educational and Psychological Measurement, 1961, 21 899-913.

Connolly, J.A. & Wantman, M.J. An exploration of oral reasoning processes in responding to objective test items. Journal of Educational Measurement, 1964,1, 59-64.

Cronbach, L.J. Essentials of psychological testing. (2nd ed.) New York Harper, 1960.

Diamond, J.J. & Williams, D.V. An identification of the process response to Taxonomy-type test items. Paper presented at meetings of the American Educational Research Association 1972.

Diamond, J.J. & Williams, D.V. Demand characteristics of multiple-choice test items. Paper presented at meetings of the National Council for Measurement in Education, 1975.

Johnson, E.S. An information-processing model of one kind of problem solving. Psychological Monographs, 1967, 78.

Kaiser, H.F. Scaling a simplex. Psychometrika, 1962, 27 155-162.

Kropp, R.P. The relationship between process and content item-response. Journal of Educational Research, 1956, 49, 385-388

Kropp, R.P., Stoker, H.W., & Bashaw, W.L. The construction and validation of tests of the cognitive processes as described in the Taxonomy of Educational Objectives., USOE, CRP No 2117, 1966. (a)

Kropp, R.P., Stoker, H.W., & Bashaw, W.L. The validation of the Taxonmy of Educational Objectives. Journal of Experimental Education, 1966, 34, 69-76. (b)

Little, E.B. & D'Asaro, G.E. Inside a student's mind during a final examination: Thoughts on wrong responses only. Peabody Journal of Education, 1956, 43, 113-119.

Miller, G.A., Galanter, E. & Pribram, K.H. Plans and the structure of behavior. New York: Holt, 1960.

Scates, D. Complexity of test items. Journal of Educational Research, 1936, 30, 77-92.

Smedslund, J. Educational Psychology. In P.R., Farnsworth, Olga McNemar, and W. McNemar. (Eds.) Annual Review of Psychology, 1964, 15, 251-276.

Williams, D.V., Diamond, J.J.; Kreher, L. and Braucher, J. Memory vs. Inference: A preliminary study of process-refernced test items. Paper presented at meetings of the American Educational Research Association, 1975.