DOCUMENT RESUME

ED 111 846 TM 004 813

AUTHOR Diamond, James J.; Williams, David V.
TITLE An Identification of the Process Response to Taxonomy-Type Test Items.
PUB DATE [Apr 72]
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1972)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS Cognitive Processes; Educational Objectives; High School Students; *Item Analysis; Problem Solving; *Response Style (Tests); *Taxonomy; *Test Construction; *Testing Problems

ABSTRACT

Although methods for estimating item difficulty are abundant, little attention has been given to the psychological processes involved when a student responds to a single test item. Mastery of educational objectives is not proven when a student supplies the correct answer to items intended to test these objectives. The student's problem solving method may differ from that intended by the test writer; there is a difference between the student employing the desired process and producing the desired product. The usefulness of the Taxonomy of Educational Objectives as a guide for item writing is questioned. (BJG)

AN IDENTIFICATION OF THE PROCESS RESPONSE TO TAXONOMY-TYPE TEST ITEMS[1]

James J. Diamond and David V. Williams

University of Pennsylvania Ithaca College

Within the area of psychometrics, methods for estimating item difficulty are abundant. The literature is replete with techniques for using statistical methods as a means of correcting item indices for various sources of error, such as guessing and partial information. Recently, however, certain psychologists and educators have approached the basic issues of item analysis (of which difficulty estimation is a central part) from a different point of view.

Smedslund (1964) has argued that the Anglo-American tradition is "test happy" and that we have paid almost no attention to the "psychological processes involved in the responses to single test items." Cronbach (1960) pointed out the general lack of concern with the nature of item solution processes. Thus, while there are many techniques for deriving a numerical estimate of difficulty, there is a need for experimental investigation of why certain items are more difficult than others. This sort of information is essential both in test-construction and for a complete understanding of the cognitive components of test-taking behavior.

Campbell (1961) proposed two basic classes of difficulty determinants which he labelled the "external" and "internal" factors. The former "influence the percent of subjects passing an item and yet are not relevant to the process(es) that the item is intended to measure (p. 901)." The latter class includes those factors which do pertain to the process the item is intended to measure.

Almost exclusively, the emphasis in research has been upon the relationship between difficulty and external factors such as item position, directions, examples, etc. This is true even though the variance in item difficulty indices is a function of both "external" and "internal" determinants.

One of the internal determinants proposed by Campbell was the effect on difficulty of changes in item complexity. (This dimension refers to the complexity of the cognitive process the item measures rather than the typographical "complexity" of the printed item e.g. the number of alternatives, length of stem).

While the area of item complexity has been studied for many years, (Scates, 1936), but it is still not well understood. More recently however, Bloom, et al., (1956) formulated the Taxonomy of Educational Objectives which has offered a method of classifying the objectives of education and thus test items used to assess mastery of these objectives.

---

1 Paper presented at Meeting of the American Educational Research Association, 1972

Bloom hypothesizes that the "cognitive domain" is divided into six broad areas: "knowledge, comprehension, application, analysis, synthesis and evaluation." The hierarchy is hypothesized to be one of cumulative complexity; that is, the behaviors within any one category include all those behaviors in the; theoretically, less complex categories. In addition, these behaviors are only expressed as those that a given test item is intended to elicit. An item is therefore classified in category i if it is intended to elicit those behaviors common to category i and to all subcategories. These authors indicated that any possible relationship between actual and intended behaviors was within the province of evaluation. This approach to the problem is not at issue. However, in order to evaluate performance in this framework one must have some degree of assurance by way of empirical relationships that the behavior students employed to solve various test problems were those intended by the test writer. The authors of the Taxonomy however appear to have simply equated mastery of objectives with supplying the correct answer to items intended to test those objectives. This implies an identity of intended and actual student behaviors which should instead be demonstrated if these approaches are to be useful in test construction. The purpose of these studies is to identify the solution strategies students employ when answering items that have been classified on the basis of intended strategies. If one cannot demonstrate a relationship between actual and intended item solution behaviors, the use of the Taxonomy in test construction must be carefully reconsidered.

METHOD

Description of Testing Materials

The items and reading passage employed in this study were adapted from those used in a study by Kropp et al., (1966a). These authors constructed a series of taxonomy-type tests and administered them to students in a series of Florida high schools. Although their tests covered four content areas, only the materials on glaciers were used in the present study.

From the items used by Kropp, et al., 50 four-choice multiple choice items were chosen to form four subtests of approximately 12 items each. As in Kropp, et al. study, only the first four levels of the Taxonomy were considered. In addition, Kropp et al. classified an item in category i on the basis of the opinion of experts in item writing and the axonomy. (See pp. 76-79 for a full description of their procedure.) Thus, the items were initially classified only on the basis of intended item-answering behaviors.

The reading passage was that used by Kropp, et al., with slight modifications. The passage contained approximately 700 words and was judged to be of high interest value for secondary school students. The vocabulary used in the test items and reading passages was approximately ninth grade level.

Subjects

The 71 subjects in the study were 11th grade high school students in suburban upstate New York. The students were generally "college-oriented" and above average in intelligence.

Procedure

Approximately one week before testing, the reading materials were distributed to the students. They were instructed to read the material over and to study in preparation for a quiz the following week.

At the testing session, the students received test booklets of 25 pages with approximately two items on each page. A space was left under each item for the student to record his method of solution immediately after answering each item. The students were instructed in doing this task to try to write down why they selected the particular answer they did. The instructions were written so that they did not suggest reasons to the students. The data generating procedure in this study was a modified "think aloud" strategy -- one which has been employed as a vehicle for assessing problem solving styles (Bloom and Broder, 1950; Miller, Galanter, and Pribram, 1960; Johnson, 1964.

The subjects were informed that they were part of a study in which it was hoped that their teachers could derive information which would help them in the classroom. The students' cooperation was requested and in the opinion of those who viewed the situation, it was generally obtained.

The recoding of the solution processes for each item was the technique used to identify the "process response" for each student. This term was borrowed from Kropp et al (1966 b). As they indicated:

> "The choice of the proper response measure is crucial if one wishes to obtain the best evidence on which to validate any behavioral measure. In the case of the Taxonomy, two possible response measures come immediately to mind. One is whether the desired intellectual process is used by the student. The other is whether the student gives a correct response to an item. The former will be referred to as the process response; the latter, the product response (p. 70)."

Preparation of Data and Analysis

An ordinal scale was constructed to serve as a set of standards against which the recorded solution processes of each student could be judged.

A level 1 response indicated agreement between the actual and intended solution processes such that the student described his solution method using behavior specified by the Taxonomy or behavior synonymous with those in the Taxonomy.

For example, the items assessing Knowledge objectives involve the process of remembering. This includes the recall of information and also the reorganization of the stimuli the item presents in order to provide cues for recall. As the Knowledge category is outlined in the Taxonomy, the material in the test items relates specifically to information which has been made available to the student either in lecture, textbook, or some other communication format. Examples of a Level 1 response for a Knowledge item would be similar to the following: "It was in the paragraph you gave us" or "I remembered it from the handout."

Preparation of Data and Analysis

An ordinal scale was constructed to serve as a set of standards against which the recorded solution processes of each student could be judged.

A level 1 response indicated agreement between the actual and intended solution processes such that the student described his solution method using behavior specified by the Taxonomy or behavior synonymous with those in the Taxonomy.

For example, the items assessing Knowledge objectives involve the process of remembering. This includes the recall of information and also the reorganization of the stimuli the item presents in order to provide cues for recall. As the Knowledge category is outlined in the Taxonomy, the material in the test items relates specifically to information which has been made available to the student either in lecture, textbook, or some other communication format. Examples of a Level 1 response for a Knowledge item would be similar to the following: "It was in the paragraph you gave us" or "I remembered it from the handout."

The second order responses involve what is sometimes called "poverty of content." That is, a general response which indicates something close to Level 1 but which lacks a particularly crucial element. For example, relative to a Knowledge item, a response such as "I thought about it" would be a Level 2 response. It is close to Level 1 but it lacks a specific reference to the handout materials.

Finally, a third order response was characterized by vague generalities. This rank also included all other responses which indicated that the students solved the problem by using a process other than the intended one. For example, this category included responses indicating that an answer to a particular item was simply recalled, when that item was classified in a category other than Knowledge. In addition, all responses which indicated that the students guessed were classified as Level 3.

It should be noted that it is possible for a student to give a Level 1 response and still have answered the item incorrectly. That is, a student can employ behaviors appropriate to a given Taxonomic level incorrectly. This indicates the distinction between an item which elicits a particular set of behaviors relative to an objective and the student's attainment of that objective as indicated by a correct response.

There are in fact four possibilities to be considered. The first has been indicated. That is, the item elicits the required behaviors but the student responds incorrectly. Secondly, it is possible for an item to elicit the required behaviors and to be answered correctly. In both of these instances, information is obtained relative to a student's attainment of a particular educational objective.

The remaining two possibilities are when an item does not elicit the behaviors relevant to a given objective and the student responds either correctly or incorrectly. Within the present framework, these latter two instances would indicated by a Level 3 response. Therefore, as part of the analysis of actual vs. intended solution processes, all responses must be considered; whether the student responded correctly or incorrectly.

For each item a distribution was made indicating the number of Level 1, Level 2, and Level 3 responses. This distribution was the result of the classification of the students' written responses by two independent raters. These raters were both given a set of standards indicating examples of Level 1, Level 2 and Level 3 responses for each of the four Taxonomic categories considered. These standards were written on the basis of the behavioral descriptions for each level included in the Taxonomy.

For any of the responses for which the judges disagreed, the higher of the two ranks was assigned. Otherwise, the ranks remained as reported by the judges.

After the distribution of rating levels was obtained for each item, the median of the ratings of the students' responses was determined. If the median was less than or equal to two, this was taken as a definition of agreement between the taxonomic process intended by the item writer and the actual process employed by the students as indicated by their written solutions. If the median was greater than two, then this item was defined as misclassified since a sufficient correspondence between actual and intended behaviors was not demonstrated.

RESULTS

Interjudge agreement averaged 76% over all subtests. That is, for all items, the judges agreed on the level of the written solution responses 76 per cent of the time. Across the four subtests from Knowledge to Analysis the figures on agreement were 79%, 87%, 86%, and 58% respectively.

It was anticipated that agreement between intended and actual student behaviors would be obtained more often for the Knowledge items than for any of the others. The reasons for this were that the Knowledge level is discussed in greatest detail in the Taxonomy, it is easiest to specify and rate ("I remembered it from the reading passage."), and item writers have had the most practice writing recall items.

The anticipated results were obtained, although they were confined to only six of the 50 items. For these items, there was the correspondence which would be expected on the basis of the use of the Taxonomy. That is, for only these items was the median of the distribution of ratings of the written solution responses less than or equal to two. Four of these were Knowledge items. In addition, process agreement was obtained for only one Application item and one Analysis item.

When this study was originally developed, it was intended that the simplex model be applied to the data. More specifically, the method of scaling a simplex, devised by Kaiser (1962) would be applied - giving that order of subtests most closely forming a scale of complexity. Then on the basis of the ratings of the written solution responses, the items would be reclassified and the Kaiser scaling recalculated. Since the reclassified subtests would be theoretically more homogeneous with respect to the process elicited, the intercorrelation matrix of subtest scores should then more closely form a perfect simplex.

The first simplex scaling was carried out and the "best order" of subtests was Knowledge, Application, Analysis, Comprehension - an order somewhat different from that posited in the Taxonomy. However, the ratings of the solution strategies did not provide data sufficient for reclassification. For those 44 items where agreement between actual and intended strategies was not found, there was no consensus as to any other taxonomy strategy. Rather a variety of response styles was evident. Students reported guessing, use of previous knowledge, partial information, and a plethora of other strategies, none of which fit the Taxonomy's frame of reference.

Therefore, by way of summary, the overall result of this study was a lack of correspondence between the actual and intended solution processes as evidenced by the students written solution strategies.

DISCUSSION

To put the findings into some perspective, it is important to note that the items and passage selected were carefully written, edited, and judged by experts to be appropriate -- a procedure far in excess of that available to a classroom teacher. Nevertheless, it is presumably desirable for the item writer to be able to predict the process an item will elicit. In fact for criterion-referenced tests, it is essential.

As indicated, agreement between actual and intended processes was not demonstrated for a majority of the items. The results of this study call into serious question the usefulness of the Taxonomy in its present form as a guide for item writing. If experts in psychometrics cannot employ these suggestions as an aid in test construction, then its value to the teacher in classroom testing is certainly questionable. Further research is needed to determine those conditions under which the actual-intended discrepancy can be reduced so that taxonomies can be made useful in test construction.

If teachers and administrators believe it is paramount that one assess cognitive processes other than simple recall, then the task for those constructing tests is great indeed. Further studies which would make it easier for students to "think aloud," involve more extended verbal reports of solution strategies, and a variety of content areas are anticipated so that those working with tests can obtain greater insights into students' probelm solving strategies.

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.