AN EMERGING TREND IN ACHIEVEMENT TESTING[1]

H. W. Stoker
Florida State University

When, at the turn of the century, Rice (4) constructed his tests for his investigation which led to the publication "The Futility of the Spelling Grind," he established the nature of educational achievement tests. His conception of educational achievement dominated education for the following half century. His idea was quite simple: measure the retention of subject matter which was taught. That his conception was widely accepted is easy to document. Mid-way through the half-century Wood (3) conducted the Commonwealth Study, a study of the quality of public schools in Pennsylvania. His prime data were "the factual knowledge which students have achieved." As recently as last year, a major test publisher issued a revised form of a classic standardized achievement test. It differs only in degree from those of Rice and Wood.

I believe that we are in a period that is drawing to a close. During this period we have depended almost completely on tests of the Rice-Wood type for the measurement of achievement. We are on the threshold of a new era in achievement testing. I would like to discuss with you what I believe to be an emerging trend in achievement testing. Perhaps time will prove it to be a dominant theme during the next half-century.

Perhaps the appropriate beginning point in a discussion about it is for you to consider a test item consisting of a stem and five alternatives. A student is directed to read the item and to choose a response. To oversimplify somewhat, his response to the item will depend on two factors; whether he is in possession of the relevant content, and whether he can bring that content to bear on the item. If he answers correctly, then we assume that he has command of the content, or he does not correctly apply the mental process, or both.

The significance of mental process in this interaction has been seriously neglected. One might ascribe the neglect as follows. First, mental processes, as such, have not been operationally defined; consequently, instruments to appraise student command of

[1] Presidential address presented at the Annual Meeting of the Florida Educational Research Association, Miami Beach, April 1964. The work reported herein was supported during 1962-63 by the Florida State University Research Council and from September 1962 to the present is supported by U. S. Office of Health, Education and Welfare, Contract OE-W-10-019, Project Number 2117.

them could not be built. Second, undue emphasis on subject matter, content achievement, has caused many to lose sight of the interwoven process variables.

For the past two years a group of us at Florida State University has been investigating the possibility of assessing students' mastery of mental process variables. The mode of investigation has been to hold accessibility of content constant over pupils then to analyze and assess their ability to manipulate the content. The operational definitions of mental processes we are using appear in the Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. (1)

The Taxonomy grew out of a series of conferences which were held during the period 1949-1953 to explore methods of classifying educational objectives. The participants were college and university examiners who were in search of a theoretical framework which would facilitate communication among them and which would serve as a basis for examining the relationship between teaching and testing. The group planned a complete taxonomy of educational objectives in three parts: the cognitive, the affective and the psychomotor domains. Considerable effort on the part of the conference participants led to the publication, in 1956, of The Taxonomy of Educational Objectives, Handbook I: Cognitive Domain.

A taxonomy, according to the Thorndike-Barnhart dictionary, is a "classification, especially in relation to its principles or laws." To build a taxonomy of educational objectives then would imply the existence of principles and laws which are common to all educative processes, whether the content be English, social studies, mathematics, science, etc. Teachers claim to be teaching for "understanding, problem solving ability, comprehension or mastery." The need for a taxonomy is emphasized when we consider that the use of relatively undefined terms, such as "problem solving", by teachers of mathematics might well imply a set of mental processes quite different from those implied when teachers of physics use the term "problem solving." The development of a standard classification system such as the Taxonomy should standardize terminology and the processes implied by terms, thereby preventing semantic confusions such as that of the physics and mathematics teacher described above.

A taxonomy has other uses. It can aid the teacher to develop a better conception of educational goals in terms of behaviors and it should suggest the possibility of new educational objectives.

Finally, since a taxonomy can be used as a basis for establishing educational goals, then it should serve, too, as a classification system for the measurement of these objectives.

The authors of the Taxonomy of Educational Objectives assumed that student behaviors could be grouped into a relatively small number of classes despite differences in curricular objectives, test materials, etc. They further assumed that the classes established would be common to levels of instruction (elementary through college), schools, and curricula.

Four principles guided the construction of the *Taxonomy*. First, the major distinctions between classes should reflect the distinctions which teachers make among student behaviors. Second, the *Taxonomy* should be logically developed and internally consistent, i.e., terms should be clearly defined and used consistently throughout. Third, the *Taxonomy* should be consistent with present knowledge and understanding of psychological phenomena. Fourth, the *Taxonomy* should be, primarily, patterned for the description of every type of educational goal, with no attempt to place any ordered value on the quality of the behaviors being classified.

On the basis of these principles and assumptions, a set of six major classes was established. The six classes were ordered hierarchically according to the complexity of the behavior in each class. The categories, from least to most complex, are:

Knowledge, behaviors which emphasize the remembering of ideas, material or phenomena. It is the lowest level in the *Taxonomy*.

Comprehension, behaviors which represent an understanding of some message transmitted through some form of communication.

Application, behaviors which emphasize the use of abstractions in specific situations.

Analysis, behaviors which breakdown a communication into its several parts in such a way that the relationship between the ideas expressed are made more clear.

Synthesis, behaviors which put together elements and parts to form a novel whole.

Evaluation, behaviors which represent judgments about the value of material and methods for specific purposes.

The hierarchy of mental processes is cumulative, i.e., each level includes all preceding levels as well as a unique component. This classification system is referred to as the *Taxonomy* of the cognitive domain. It provides the theoretical framework for our research on the measurement of educational achievement.

Being good researchers, or perhaps regressive ones, we decided to center initial studies on the validity of the Taxonomy, the framework we intended to use in our proposed studies of the measurement of achievement. In short, the first task was validation of the Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. Two specific questions were considered: Can judges agree on the assignment of items to Taxonomy categories? Can the imputed hierarchical structure of the Taxonomy be supported by empirical evidence?

The first question, can judges agree on the assignment of items to the Taxonomy categories, was answered affirmatively in two separate studies. One study dealt with published, standardized tests; the other with specially constructed tests.

Two well-known, standardized tests were selected for analysis, one a test of reading comprehension, the other a test of arithmetic computation. A panel of judges, all familiar with the Taxonomy, were asked to classify each item according to the mental process (knowledge, comprehension, application, analysis, synthesis, and evaluation) which they judged would be required in responding to the item. While the agreement among the judges was not unanimous for all items, there was sufficient agreement to attest to the validity of the Taxonomy as a criterion of classification.

Of particular interest, although not the central purpose of this phase of the research, is the distribution of items in these tests with respect to Taxonomy categories. When the model classifications of the items were considered 93% of the items in the reading comprehension test were classified as either Knowledge or Comprehension. For the arithmetic computation test, Knowledge, Comprehension and Application accounted for approximately 75% of the item classifications. These results testify to the emphasis on content rather than process in the "typical" standardized test.

The second part of the answer to this question was based on items which were written to conform to the operational definitions of behaviors contained in the Taxonomy. The problem in this context was, if item writer "A" writes an item designed to evoke behavior "S," will judges "B," "C," "D," etc., classify the item as measuring "S." Two sets of items were constructed, one set based on a science reading passage and the other set based on social studies content.

Five judges classified the science items. Eleven of thirty-six items were unanimously classified as congruent with the categories for which the items were written. On nine other items, only one judge deviated in his classifications from those of the other four. (Six of the nine disagreements were attributed to one judge.) On all but two of the remaining sixteen items, three of five judges classified each item congruently with the process it was intended to evoke.

Four judges classified each of 39 items based on social studies content. For eleven items their agreement was unanimous and perfectly related to the process category the items were intended to evoke; for sixteen items, three of four judges agreed with each other and the intended category. For each of the remaining items two judges were in agreement.

In summary, raters do tend to agree among themselves as to the behaviors required and do tend to classify items congruent with the behaviors the items were intended to evoke.

The second question dealt with the empirical validation of the hierarchical structure of the Taxonomy. If the categories are hierarchical, from Knowledge through Evaluation, and also cumulative, i.e., any given category includes the behaviors of all lower-order categories, then data derived from tests constructed to appraise all of the Taxonomy behaviors should exhibit the following characteristics. First, the mean scores for items in each category should decrease as the complexity of the category increases. Second, analysis of the inter-correlation matrix of category scores should reveal a simplical structure as described by the Guttman Radex Theories. (2)

The experimental tests, consisting of items constructed to measure the Taxonomy behaviors, were administered to approximately 1,000 students in grades 9-12 from two schools. The hypothesized order of mean category scores was supported; mean process scores did decrease as the level of complexity increased. This not only supports the hierarchical structure of the Taxonomy, but also indicates some construct validity for the experimental tests. A perfect simplex, according to Guttman's theories, requires that certain partial correlations vanish. In particular, $r_{ik \cdot j} = 0$ for $i < j < k$. This condition will obtain only if,

$$r_{ik} = r_{ij}\, r_{jk}, \text{ where } i < j < k.$$

While perfect simplical structure was lacking in the correlation matrices, there was a definite trend toward a quasi-simplex, a modification of the perfect simplex (inasmuch as the requirements for the perfect simplex are extremely rigorous, it was not surprising that the conditions were not met).

Referring back to our two questions, we might say that judges can agree on the assignment of items to Taxonomy categories and there is general support for the imputed hierarchical structure of the Taxonomy.

One or two other findings are perhaps worth mentioning. The experimental tests were administered to students in grades 9-12;

for each category, there was a general increase in mean performance as the grade level increased from 9-12. This may well indicate that the processes are perhaps general, learned outcomes of the educative processes. The experimental tests were constructed in the form of the typical reading comprehension test and, hence, these differences may reflect differences in reading ability rather than maturation level of the process.

When process scores were correlated with I.Q., the correlation decreased as the level of complexity of the process increased. The most plausible explanation for this would be that the I.Q. test measures only those processes at the lower end of the hierarchy. The pattern of correlations with I.Q., coupled with the pattern of process mean scores, implies that the tests are measuring something other than maturation.

The measurement of achievement for the past few decades can be characterized by tests in the Rice-Wood tradition. The majority of these tests have emphasized the acquisition of content. Scores from these tests have been used successfully to identify students who did not have command of the relevant content. However, little has been done to identify the mental processes involved in the manipulation of this content.

I predict that there will be a shift in emphasis during the next few decades in the direction of a more careful examination of the mental processes involved in learning. Perhaps the test battery of the future will be content and process oriented. Scores from such a battery might tell us not only the content areas in which a student is weak but also the mental processes which must be learned to facilitate the strengthening of the weak areas.

The validation studies discussed here indicate the relatively low level of mental process involved in the typical standardized test. They also indicate the usefulness of the Taxonomy as a criteria for classification for behavior and as a set of operational definitions which can be used to determine the extent to which these behaviors have been learned. The studies reported here are just a beginning; the work has continued and its dimensions have been extended. I believe the trend toward the inclusion of process measurement to be well worth watching.

References

1. Bloom, B. S., Ed. Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. New York: Longmans, Green and Co., 1956.

2. Guttman, Louis. "A New Approach to Factor Analysis: The Radex." In Mathematical Thinking in the Social Sciences, P. F. Lazarfeld, ed. Glencoe, Illinois: The Free Press, 1954.

3. Learned, W. S. and Wood, B. D. The Student and his Knowledge. New York: Carnegie Foundation for the Advancement of Teaching, 1938.

4. Rice, J. M. "The Futility of the Spelling Grind." Forum, Vol. 23, 1897.