N086-630 ERIC REPORT RESUME

ED 010 200 1-13-67 24

THE DEVELOPMENT OF A SEQUENTIALLY SCALED ACHIEVEMENT TEST.
COX, RICHARD C. * GRAHAM, GLENN T.
TGZ65143 UNIVERSITY OF PITTSBURGH, LEARNING R AND D CTR., PA.
BR-5-0253-8
- -66
EDRS PRICE MF-$0.09 HC-$0.48 12P.

*SEQUENTIAL LEARNING, *TEST CONSTRUCTION, KINDERGARTEN, FIRST GRADE, SECOND GRADE, *INDIVIDUAL INSTRUCTION, *ACHIEVEMENT TESTS, SEQUENTIAL APPROACH, RATING SCALE, MEASUREMENT INSTRUMENTS, *BEHAVIOR RATING SCALE, ABILITY IDENTIFICATION, PITTSBURGH, PENNSYLVANIA

THIS WORKING PAPER IS CONCERNED WITH THE DEVELOPMENT OF SEQUENTIALLY SCALED TESTS FOR STUDENTS RECEIVING INDIVIDUALIZED INSTRUCTION. THESE TESTS WERE DESIGNED TO INDICATE MASTERY OF DESIRED BEHAVIORS RATHER THAN TO DISCRIMINATE AMONG INDIVIDUALS. OBJECTIVES WERE IDENTIFIED AND ARRANGED SEQUENTIALLY. TESTS WERE GIVEN TO KINDERGARTEN, FIRST-GRADE AND SECOND-GRADE CHILDREN IN ORDER TO OBTAIN A WIDE RANGE OF ABILITY LEVELS. THE RESULTS INDICATED THAT IT WAS POSSIBLE TO DEVELOP A SEQUENTIALLY SCALED ACHIEVEMENT TEST. (LB)

The Development of a Sequentially Scaled Achievement Test[1]

Richard C. Cox and Glenn T. Graham

University of Pittsburgh

A program of individualized instruction demands a re-examination of traditional testing procedures. In the typical learning situation instructional materials and rate are held constant, and achievement testing at the end of some specified unit of work is designed to rank students according to varying levels of achievement. Individualized instruction, on the other hand, allows each individual student to set his own learning pace; yet, performance criteria for successful completion of some specified unit of work are identical for all students (Coulson and Cogswell, 1965).

Items for achievement testing in the latter situation should be designed to indicate whether or not the required behaviors have been mastered; not to discriminate among individuals. Students must be compared to an absolute standard as opposed to a normative standard, the student's score reflecting the degree of his performance with that of other individuals. This distinction between norm and criterion-referenced measures has been made by Glaser (1963).

---

[1] Paper read at the annual meeting of the American Educational Research Association, Chicago, Illinois, February 1966.

The research and development reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education and Welfare under the provisions of the Cooperative Research Program.

A content-referenced measure provides considerable information for making decisions concerning student advancement. For example, a score of 80 per cent indicates that a student has successfully mastered 80 per cent of the specified behaviors. However, unless the test is designed to measure performance on only one behavior, the total test score does not indicate which behaviors the student has or has not mastered. In order to obtain this information, student performance on each item must be examined.

One solution to this rather laborious task would be a test in which the total score would indicate the response pattern of the individual. Such tests have been employed in the investigation of attitudes, and have been analyzed using the Guttman (1944) "Scalogram Analysis." Essentially, the analysis includes the ranking of scores from highest to lowest, and the ranking of items from most favorable to least favorable. Theoretically, those students with the highest scores (highest being most favorable) would have answered only the most favorable items; those scoring low would have answered only the least favorable items, etc. The analysis yields a coefficient of reproducibility which indicates how well an individual's response pattern can be reproduced knowing his total score. The value of .90 was arbitrarily established as an acceptable lower limit.

Applying this technique to achievement testing would yield valuable information. If the behaviors to be tested could be arranged in a sequential order, and the test were scalable, a student who obtained a score of 5 would have answered items 1, 2, 3, 4, and 5 and no more. A student could not score 7 unless he answered 1 through 7 and did not answer any items beyond 7. Knowing the behaviors these items represent, the score on the test indicates to the teacher, guidance counselor, or researcher those behaviors the student has mastered and those behaviors he has yet to master. The present study is an attempt to develop such a test.

Procedure and Results

The first step in the development of any test is the identification of objectives to be tested. In a test designed to be scalable, the objectives must be arranged sequentially. In this study the terminal objective to be tested was the student's ability to add 2 two-digit numerals involving carrying. Using this as a starting point, the question was asked, "What skills must have been mastered previously in order to master this objective?" With this question as a guide, the list of fifteen objectives presented in Figure I was developed.

Figure 1

Objectives	Sample Test Items
The student is able to:	
1. Recognize numerals from 1 to 10.	1. 1 2 3 4 "Draw a circle around the 2."
2. a) Determine which numeral comes before or after another numeral.	2. a) 10 8 5 2 "Draw a circle around the number that comes just after 7."
b) Determine which of 2 numerals is the largest or smallest.	b) 7 5 "Draw a circle around the largest numeral."
3. Discriminate between +, -, =, ≠.	3. + - = ≠ "Draw a circle around the sign which means to add."
4. a) Add two single-digit numerals with sums to 10, vertically.	4. a) 5 +4
b) Add two single-digit numerals with sums to 10, horizontally.	b) 3 + 1 =
5. a) Add two single-digit numerals involving carrying, horizontally.	5. a) 8 + 3 =
b) Add two single-digit numerals involving carrying, vertically.	b) 7 +8

Figure I cont.

Objectives	Sample Test Items
6. a) Add three single-digit numerals involving carrying, vertically.	6. a) 1 / 8 / +4
b) Add three single-digit numerals involving carrying, horizontally.	b) 7 + 3 + 2 =
7. a) Place one and two-digit numerals in a column so they could be added.	7. a) 15 16 2 "Place these numerals in a column so they could be added."
b) Determine which column of numerals is written so they could be added.	b) 15 / 16 / 2 ; 15 / 16 / 2 ; 15 / 16 / 2 "Draw a circle around the column which is written so it could be added.
8. Add 2 two-digit numerals without carrying.	8. 20 / +11
9. Add 2 three-digit numerals without carrying.	9. 215 / +723
10. Add 2 two-digit numerals with carrying.	10. 58 / +36

With the exception of one task for objective 7b, there were from 2 to 5 tasks constructed for each of the objectives. The total number of tasks for the entire test was 50. The tasks pertaining to each objective were combined to form one "item" for each objective. This procedure is similar to the "H-technique" suggested by Stouffer, Borgatta, Hays, and Henry (1953). As an example, consider the three tasks:

20	36	54
+11	+42	+33

These tasks would compose one "item" testing objective 8.

This procedure was followed, and a 15 item test was constructed. The test was administered to a kindergarten, first, and second grade in order to obtain a wide range of ability levels. The possible total score range was from 0 to 15. Students were ranked according to this total score and the response pattern was plotted. This pattern indicated that some of the items were not in the correct position to obtain the maximum coefficient of reproducibility, i.e., the postulated sequence of objectives was not empirically verified. The items were rearranged in order to yield the maximum reproducibility coefficient. The response pattern obtained after the items had been rearranged yielded a reproducibility coefficient of .961.

As a further revision, objectives 3, 7a, and 7b and their corresponding items were omitted—objective 3 because it was dependent on a specific curriculum, and objectives 7a and 7b because of ambiguous directions. The final arrangement of items yielded a reproducibility coefficient of .977.

According to Guttman in Stouffer (1950), the coefficient of reproducibility is a necessary but not a sufficient criterion for scalability. Since the reproducibility coefficient for a given item cannot be less than the proportion of responses occurring in the most frequently chosen category, Guttman suggests that too many extreme items, 80 per cent or greater in any one category can spuriously raise the reproducibility coefficient. Herbert Menzel (1953) suggests a procedure, which, when taken in conjunction with the reproducibility coefficient, further contributes evidence of scalability. Menzel suggests a coefficient of scalability which determines the degree to which the individual's performance can be reproduced from knowledge of the marginal totals. The coefficient prevents one from spuriously attributing high scalability from a sample composed of many extreme items and/or individuals. A coefficient of .65 or better is established as a criterion. The scalability coefficient for the revised test was .902.

Although the revised test met the criteria for scalability, it had never been administered in its present form. As a validation study, the final revision was administered to a different kindergarten, first, and second grade. This new response pattern yielded a reproducibility coefficient of .970 and a scalability coefficient of .792.

Discussion

The results indicate that it is indeed possible to develop a sequentially scaled achievement test. However, these results must be tempered by the fact that the test is based upon a restricted

area of subject matter. At the present time additional tests are being developed in other areas of mathematics covering a wider range of objectives. These areas include subtraction, addition, time telling, numeration, and money. The reproducibility coefficients obtained range from .85 to .96 on the initial test administration. Further replications with more complex skills and with larger and more heterogeneous samples would be desirable.

The results of the study also should be tempered with the realization that the item responses may be a function of prior educational experiences, in school or elsewhere, to which the students have been exposed. This is not to say, however, that it is impossible to have certain skills which are necessarily prerequisite to others but rather to suggest a possible contaminating factor. It also seems reasonable to hypothesize that by manipulating the content taught in the classroom one could dictate a series of objectives which would yield an empirically scaled test.

One obvious result of the study is that the logical ordering of objectives is not sufficient for the establishment of a scalable test. Empirical evidence must be obtained to verify or refute the postulated order.

References

Coulson, J. E. and Cogswell, J. F. Effects of individualized instruction on testing. Journal of Educational Measurement, 1965, 2, 59-64.

Glaser, R. Instructional technology and the measurement of learning outcomes: some questions. American Psychologist, 1963, 18, 519-521.

Guttman, L. A basis for scaling qualitative data. The American Sociological Review, 1944, 9, 139-150.

Menzel, H. A new coefficient for scalogram analysis. Public Opinion Quarterly, 1953, 17, 268-280.

Stouffer, et. al. Studies in social psychology in World War II, Vol. 4, Measurement and Prediction. Princeton: Princeton University Press, 1950.

Stouffer, S. A., Borgatta, E. F., Hays, D. G., Henry, A. F. A technique for improving cumulative scales. Public Opinion Quarterly, 1952, 16, 273-291.