ON REPORTING STUDENT ACHIEVEMENT: THE NEED FOR MEANINGFUL TEST RESULTS

by Kathryn W. Linden and Wayne M. Garrison

Purdue University

The purpose of this paper is to provide users of teacher-made tests with a computer program[1] designed to improve the reporting of student performance on academic tasks. It is believed that greater specificity in what is measured and reported to students serves to clarify more accurately the diagnostic function of testing and, concomitantly, to direct attention to the complex nature of achievement in schools and colleges. The method of reporting test results presented herein was adapted from the traditional concept of a table of specifications used to combine specific subject matter with behavioral expressions of desired student outcomes.

RATIONALE

Although the sequence of steps to be followed in the preparation of classroom tests may vary from one teacher to another, or even from one textbook to another, it is agreed generally that achievement tests should be designed to reflect (1) the instructional emphases or content areas covered during instruction and (2) the types of cognitive outcomes students are expected to demonstrate at the end of instruction. In order to relate instructional objectives to specific course content, test constructors frequently utilize a two-dimensional organizational scheme, variously termed a table of specifications, specification chart or test blueprint, in order to obtain a reasonable degree of correspondence between what was taught and what is to be tested (see, for example, Ebel, 1972; Gronlund, 1976; Kryspin & Feldhusen, 1974; Thorndike, 1971). If a specification chart which relates instructional content and cognitive outcomes can be used to guide the teacher in the development and construction of test items, it can be used also as a guide to providing students with

feedback concerning performance on tasks classified in two dimensions. An example of a table of specifications is presented as Table 1.
1. Copyright 1976 Purdue Research Foundation.

Discussing the use of single and multiple marks to describe performance in a classroom situation, Ebel (1972) questioned whether the traditional single letter, or a number score, "does justice" to the complex aspects of achievement. Recognizing that the goals of education include diverse processes such as the communication of knowledge, clutivation of understanding, development of skills and abilities, encouragement of interests and exemplification of ideals, a single mark, or a single score, on an achievement test ultimately results in the loss of important information to both student and teacher. Consequently, the diagnostic and self- evaluative functions of testing which serve to identify student strengths and weaknesses and to indicate to the teacher which instructional areas may be in need of improvement frequently are not realized.

Related to the problems surrounding the interpretation of total test scores, Baskin (1975) argued that traditional right-wrong scoring methods fail to consider the complex configurations of correct responses by which two or more examinees may arrive at identical total test scores. To compensate for the non-uniqueness of test scores, Baskin further proposed the use of a configuration-scoring paradigm, primarily as a technique for studying characteristics of examinees having identical raw scores. The conclusion to be drawn is, then, that identical total test scores do **not** guarantee identical evaluations of student performance on a cognitive task when scoring efforts become more analytical and descriptive.

As appealing as Baskin's (1975) scoring technique may be, most educators who use a restricted, or selected, response test format (multiple-choice, true-false, matching) probably find the use of the right-wrong scoring procedure to be more practical and efficient. Furthermore, if a teacher has taken the time to develop a measure in conformity with a specification chart, a minimum of additional effort can provide the basis for reporting test results which are meaningful and useful to examinees.

The proposed method for reporting meaningful test results is addressed primarily to educators who use a selected-response type of test format as part of their classroom instruction and those for whom computer facilities are available for processing and analyzing test data. The method involves a basic FORTRAN program (TESTRPT) which produces individual descriptions of student test performance. Both criterion-referenced data (percentage correct) and norm-referenced data (T-scores) are provided in order to aid the evaluative aspect of testing. Item statistics are not provided insofar as a number of commercially available programs are available to meet this need (e.g., Dixon, 1968; Nie, Bent, & Hull, 1970; Veldman, 1967). Rather, the present program focuses on providing information

Table 1
A Table of Specifications For An Examination
On Test Planning And Construction

Topic Content / Cognitive Outcome	Knowledge of Facts, Concepts & Principles	Comprehension of Concepts & Principles	Application of Concepts & Principles	Analysis of Concepts & Principles	Synthesis of Concepts & Principles	Evaluation of Concepts & Principles	Total
Functions of Measurement	2	3	3	2	1	1	12 (20%)
Behavioral Objectives	2	3	3	2	1	1	13 (22%)
Item Types	2	3	3	2	1	1	11 (18%)
Item Construction	2	3	3	2	2	2	14 (23%)
Administration & Scoring	2	3	3	2	0	0	10 (17%)
Total	10 (17%)	15 (25%)	15 (25%)	10 (17%)	5 (8%)	5 (8%)	60 (100%)

relevant to test performance at a level which may serve the needs of the test-taker to a greater extent than do other methods of reporting achievement.

TECHNIQUE

The TESTRPT computer program was developed for use with classroom achievement tests employing objective-type item formats (e.g., m-c, t-f, matching, etc.). However, any type of test data may be analyzed by the TESTRPT program provided that (1) the data are quantifiable; (2) performance can be scored right-wrong or yes-no; and (3) the instrument has been constructed, or can be described, in a way that is congruent with a two-way table of specifications.

In order to utilize the present scoring program, response data must appear in computer punched form. This limitation requires that individual score sheets, or score cards, administered in optical scan format must be converted to punched format by means of an editing procedure. Input to the program consists of the responses of **N** individuals to each of **k** items composing the test. At present, TESTRPT has the capability of handling 150 individuals responding to a maximum of 150 items.

Responses to items are matched against a score key and are coded 1 for a correct response and 0 for an incorrect response. Two additional keys must be supplied in order to identify item groupings according to taxonomic classification and content areas. In addition, title information (test identification, course number, semester, etc.) and descriptions of content areas and taxonomic levels utilized are required. An option for multiple keying of correct responses to a single item does **not** exist. An illustration of the format for the output generated by TESTRPT is presented in Table 2.

This method of reporting test results provides the basis for both criterion-referenced interpretation (percentage correct) and norm-referenced interpretation (standard T-score) of individual student performance. Moreover, test items are classified according to specific content and taxonomic level, thus providing a useful description of individual strengths and weaknesses for a specified set of academic tasks. The individual computerized report identifies test items which were answered incorrectly by the examinee in order to encourage further study in those areas presenting particular problems for a given student. Finally, total group results are included in the form of summary test statistics.

The value of such a reporting system lies in its ability to provide descriptive feedback to the individual test-taker regarding test performance in a manner which optimizes the self-evaluative function of testing. Moreover, a reporting scheme of this type may serve to communicate information to the teacher regarding the effectiveness of instructional strategies.

An Illustration of the Individualized Form for Reporting Student Test Performance

STUDENT ID ... 00001 COURSE ... ED 524

EXAMINATION ... TEST PLANNING AND CONSTRUCTION SEMESTER ... SUMMER, 1975

ITEMS CLASSIFIED ACCORDING TO TAXONOMIC LEVEL				
A. KNOWLEDGE	(10.)*	8.	80.0	61.
B. COMPREHENSION	(15.)	12.	80.0	62.
C. APPLICATION	(15.)	11.	73.3	59.
D. ANALYSIS	(10.)	7.	70.0	54.
E. SYNTHESIS	(5.)	3.	60.0	50.
F. EVALUATION	(5.)	4.	80.0	65.

ITEMS CLASSIFIED ACCORDING TO CONTENT AREA		NO. CORRECT	PCT. CORRECT	T-SCORE**
A. FUNCTIONS OF MEASMT	(12.)	10.	83.3	63.
B. BEHAVIORAL OBJECTIVES	(13.)	10.	76.9	58.
C. ITEM TYPES	(11.)	4.	36.4	45.
D. ITEM CONSTRUCTION	(14.)	11.	78.6	57.
E. ADMIN & SCORING	(10.)	10.	100.0	62.
TOTAL	(60.)	45.	75.0	65.

* NUMBERS APPEARING IN PARENTHESES REPRESENT THE TOTAL NUMBER OF ITEMS INCLUDED IN A PARTICULAR CATEGORIZATION.

** T-SCORES ARE COMPUTED UTILIZING A MEAN OF 50 AND A STANDARD DEVIATION OF 10. BASED UPON TOTAL TEST SCORE DATA FOR THIS EXAMINATION, THE FOLLOWING STATISTICS WERE COMPUTED:

MEAN = 36.3, STANDARD DEVIATION = 5.82, KR 20 RELIABILITY COEFFICIENT = .69

STANDARD ERROR OF MEASUREMENT = 3.24

PLEASE CHECK YOUR RESPONSES TO THE FOLLOWING ITEMS SINCE YOU APPARENTLY MADE SOME MISTAKES.

ITEM	TAXONOMIC LEVEL	CONTENT AREA	YOUR RESPONSE	KEYED RESPONSE
4	KNOW	A	1	3
5	KNOW	A	3	4
14	COMP	B	3	2
16	APP	B	2	1
22	ANAL	B	4	3
27	APP	C	1	2
28	ANAL	C	2	3
29	APP	C	2	4
31	COMP	C	4	1
32	ANAL	C	1	2
34	EVAL	C	3	4
36	SYN	C	3	1
40	SYN	D	4	2
47	APP	D	1	3
49	COMP	D	2	1

INSTRUCTOR COMMENTS . . .

While the TESTRPT program bears general resemblance to the computer-assisted instruction (CAI) movement in education, it has the advantage of being easier to operate and requires little previous exposure to computer technology. Instituitions with computer facilities capable of accommodating statistical packages such as BMD (Dixon, 1968), SPSS (Nie et al., 1970) and EDSTAT (Veldman, 1967) should not experience difficulty in field length requirements for the TESTRPT program. It utilizes standard FORTRAN language and includes a sufficient number of comment cards to permit individuals with limited programming knowledge to process data efficiently. TESTRPT is available presently in card form. Documentation, including a sample set of data, also is provided with each request.

CONCLUSIONS

Because there is need to report student performance on classroom measures in a manner consistent with the diagnostic and self-evaluative functions of testing, the use of a reporting system which employs only single letter grades or total test scores may be of questionable value. However, if a test constructor has developed measures in congruence with a table of specifications, the problem of scoring data in an analytical and descriptive manner may be reduced to a simplified computer task. The program described in this paper was developed in order to provide test-takers with performance information thought to be more meaningful than that provided by a single test score.

Inquiries should be addressed to: Dr. Samuel Mark, Purdue Research Foundation, Purdue University, West Lafayette, Indiana 47907.

References

Baskin, D. A configuration-scoring paradigm for identical raw scores. **Journal of Educational Measurement,** 1975, **12,** 3-5.

Dixon, W. J. (Ed.) **BMD: Biomedical computer programs.** Berkeley: University of California Press, 1968.

Ebel, R. L. **Essentials of educational measurement** (2nd Ed.). Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Garrison, W. M. & Linden, K. W. **Individualized Test Score Program.** West Lafayette, Indiana, Purdue Research Foundation, 1976.

Gronlund, N. E. **Measurement and evaluation in testing** (3rd Ed.). New York: MacMillan, 1976.

Kryspin, W. J. & Feldhusen, J. F. **Developing classroom tests: A guide for writing and evaluating test items.** Minneapolis, Minnesota: Burgess, 1974.

Nie, N. H., Bent, D. H., & Hull, C. H. **Statistical package for the social sciences.** New York: McGraw-Hill, 1970.

Thorndike, R. L. (Ed.) **Educational measurement** (2nd Ed.). Washington, D.C.: American Council on Education, 1971.

Veldman, D. J. **Fortran programming for the behavioral sciences.** New York: Holt, Rinehart, and Winston, 1967.