INSTRUCTIONAL TECHNOLOGY AND THE MEASUREMENT OF LEARNING OUTCOMES:

SOME QUESTIONS [1]

ROBERT GLASER

University of Pittsburgh

EVALUATION of the effectiveness of teaching machines and programed learning, and of broadly conceived instructional systems, has raised into prominence a number of questions concerning the nature and properties of measures of student achievement. In the evaluation of instructional systems, the attainment of subject matter knowledge and skill as well as other behavioral outcomes must, of course, be considered, but the remarks in this paper will be restricted primarily to the measurement of subject matter proficiency, as it may be defined by recognized subject matter scholars.

Achievement measurement can be defined as the assessment of terminal or criterion behavior; this involves the determination of the characteristics of student performance with respect to specified standards. Achievement measurement is distinguished from aptitude measurement in that the instruments used to assess achievement are specifically concerned with the characteristics and properties of present performance, with emphasis on the meaningfulness of its content. In contrast, aptitude measures derive their meaning from a demonstrated relationship between present performance and the future attainment of specified knowledge and skill. In certain circumstances, of course, this contrast is not quite so clear, for example, when achievement measures are used as predictor variables.

The scores obtained from an achievement test provide primarily two kinds of information. One is the degree to which the student has attained criterion performance, for example, whether he can satisfactorily prepare an experimental report, or solve certain kinds of word problems in arithmetic. The second type of information that an achievement test score provides is the relative ordering of individuals with respect to their test performance, for example, whether Student A can solve his problems more quickly than Student B. The principal difference between these two kinds of information lies in the standard used as a reference. What I shall call criterion-referenced measures depend upon an absolute standard of quality, while what I term norm-referenced measures depend upon a relative standard. Distinctions between these two kinds of measures have been made previously by others (Flanagan, 1951; Ebel, 1962).

Criterion-Referenced Measures

Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on this continuum as indicated by the behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum. The term "criterion," when used in this way, does not necessarily refer to final end-of-course behavior. Criterion levels can be established at any point in instruction where it is necessary to obtain information as to the adequacy of an individual's performance. The point is that the specific behaviors implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of these knowledge levels. It is in this sense that measures of proficiency can be criterion-referenced.

Along such a continuum of attainment, a student's score on a criterion-referenced measure pro-

[1] Symposium address presented at meetings of American Educational Research Association, Chicago, February 1963. This paper is concerned with student educational achievement; however, similar notions have been expressed with respect to the human component in man-machine systems in R. Glaser and D. J. Klaus (1962).

vides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others.

Norm-Referenced Measures

On the other hand, achievement measures also convey information about the capability of a student compared with the capability of other students. In instances where a student's *relative* standing along the continuum of attainment is the primary purpose of measurement, reference need not be made to criterion behavior. Educational achievement examinations, for example, are administered frequently for the purpose of ordering students in a class or school, rather than for assessing their attainment of specified curriculum objectives. When such norm-referenced measures are used, a particular student's achievement is evaluated in terms of a comparison between his performance and the performance of other members of the group. Such measures need provide little or no information about the degree of proficiency exhibited by the tested behaviors in terms of what the individual can do. They tell that one student is more or less proficient than another, but do not tell how proficient either of them is with respect to the subject matter tasks involved.

In large part, achievement measures currently employed in education are norm referenced. This emphasis upon norm-referenced measures has been brought about by the preoccupation of test theory with aptitude, and with selection and prediction problems; norm-referenced measures are useful for this kind of work in correlational analysis. However, the imposition of this kind of thinking on the purposes of achievement measurement raises some question, and concern with instructional technology is forcing us toward the kind of information made available by the use of criterion-referenced measures. We need to behaviorally specify minimum levels of performance that describe the least amount of end-of-course competence the student is expected to attain, or that he needs in order to go on to the next course in a sequence. The specification of the characteristics of maximum or optimum achievement after a student has been exposed to the course of instruction poses more difficult problems of criterion delineation.

The Uses of Achievement Measurement

Consider a further point. In the context of the evaluation of instructional systems, achievement tests can be used for two principal purposes. First, performance can be assessed to provide information about the characteristics of an individual's present behavior. Second, achievement can be assessed to provide information about the conditions or instructional treatments which produce that behavior. The primary emphasis of the first use is to discriminate among individuals. Used in the second way, achievement tests are employed to discriminate among treatments, that is, among different instructional procedures by an analysis of *group* differences.

Achievement tests used to provide information about *individual* differences are constructed so as to maximize the discriminations made among people having specified backgrounds and experience. Such tests include items which maximize the likelihood of observing individual differences in performance along various task dimensions; this maximizes the variability of the distribution of scores that are obtained. In practical test construction, the variability of test scores is increased by manipulating the difficulty levels and content of the test items.

On the other hand, achievement tests used primarily to provide information about differences in treatments need to be constructed so as to maximize the discriminations made between *groups* treated differently and to minimize the differences between the individuals in any one group. Such a test will be sensitive to the differences produced by instructional conditions. For example, a test designed to demonstrate the effectiveness of instruction would be constructed so that it was generally difficult for those taking it before training and generally easy after training. The content of the test used to differentiate treatments should be maximally sensitive to the performance changes anticipated from the instructional treatments. In essence, the distinction between achievement tests used to maximize individual differences and tests used to maximize treatment or group differences is established during the selection of test items.

In constructing an achievement test to differenti-

ate among *individuals* at the end of training, it would be possible to begin by obtaining data on a large sample of items relating to curriculum objectives. Item analysis would indicate that some test items were responded to correctly only by some of the individuals in the group, while other items were answered correctly by all members of the group. These latter 1.00 difficulty level items, since they failed to differentiate among individuals, would be eliminated because their only effect would be to add a constant to every score. The items remaining would serve to discriminate among individuals and thus yield a distribution of scores that was as large as possible, considering the number and type of items used.

On the other hand, if this test were constructed for the purpose of observing *group* instead of individual differences, the selection of items would follow a different course. For example, where instruction was the treatment variable involved, it would be desirable to retain test items which were responded to correctly by all members of the posttraining group, but which were answered incorrectly by students who had not yet been trained. In a test constructed for the purpose of differentiating groups, items which indicated substantial variability within either the pre- or posttraining group would be undesirable because of the likelihood that they would cloud the effects which might be attributable to the treatment variable.

In brief, items most suitable for measuring individual differences in achievement are those which will differentiate among individuals all exposed to the same treatment variable, while items most suitable for distinguishing between groups are those which are most likely to indicate that a given amount or kind of some instructional treatment was effective. In either case, samples of test items are drawn from a population of items indicating the content of performance; the particular item samples that are drawn, however, are those most useful for the purpose of the kind of measurement being carried out. Hammock (1960) has previously discussed such a difference.

The points indicated above reflect the achievement measurement concerns that have arisen in my own work with instructional technology. There is one further point which must be mentioned, and that is the use of diagnostic achievement tests prior to an instructional course. It appears that, with the necessity for specifying the entering behavior that is required by a student prior to a programed instructional sequence, diagnostic assessment of subject matter competence must take on a more precise function. This raises the problem of developing an improved methodology for diagnostic achievement testing. In this regard, researchers using programed instructional sequences to study learning variables point out that prior testing influences learning, and that this effect must be controlled for in determining the specific contribution of programing variables. In an instructional sense, however, the influence and use of pretesting is an important variable for study since it is not the terminal criterion behavior alone which dictates required instructional manipulations, but the differences between entering and terminal behavior. Furthermore, pretesting of a special kind may contribute to "motivation" by enhancing the value of future responses; there is some indication that this may be brought about by prior familiarity with future response terms (Berlyne, 1960, pp. 296–301) or by permitting some early aided performance of the terminal behavior eventually to be engaged in (Taber, Glaser, & Schaefer, 1963, Ch. 3).

In conclusion, the general point is this. Test development has been dominated by the particular requirements of predictive, correlational aptitude test "theory." Achievement and criterion measurement has attempted frequently to cast itself in this framework. However, many of us are beginning to recognize that the problems of assessing existing levels of competence and achievement and the conditions that produce them require some additional considerations.

REFERENCES

BERLYNE, D. E. *Conflict, arousal, and curiosity.* New York: McGraw-Hill, 1960.

EBEL, R. L. Content standard test scores. *Educ. psychol. Measmt.*, 1962, **22**, 15–25.

FLANAGAN, J. C. Units, scores, and norms. In E. T. Lindquist (Ed.), *Educational measurement.* Washington, D. C.: American Council on Education, 1951. Pp. 695–763.

GLASER, R., & KLAUS, D. J. Proficiency measurement: Assessing human performance. In R. Gagné (Ed.), *Psychological principles in system development.* New York: Holt, Rinehart & Winston, 1962. Pp. 421–427.

HAMMOCK, J. Criterion measures: Instruction vs. selection research. *Amer. Psychologist,* 1960, **15**, 435. (Abstract)

TABER, J. I., GLASER, R., & SCHAEFER, H. H. *A guide to the preparation of programmed instructional materials.* Reading, Mass.: Addison-Wesley, in press.