PRINCIPLES OF EVALUATION AND MEASUREMENT

REGINALD JONES

INTRODUCTION

The purpose of this article is to introduce the series on Testing and Grading of the Ethiopian Journal of Education by discussing some of the fundamental concepts underlying the measurement and evaluating of student achievement. Articles in the series will be issued periodically.

Virtually everything people do involves an element of evaluation. Even the simplest conversation between two persons is accompanied by the attempts of each to determine his impact upon the other. Although evaluation is not at all unique to education, its application is somewhat more rigorous in this area than in many others. The majority of students are, at least in part, motivated to learn by the promise of a diploma or degree and by the fact that grades are assigned in individual courses. The frenzy of activity on most campuses during the final examination period is itself evidence of the fact that evaluation is part and parcel of the educational process. For most students, school and tests are almost synonymous, and "fair" grading is one of the marks of a good teacher.

MEASURING AND EVALUATING ACHIEVEMENT

The measurement of physical properties is commonoplace in daily life. We make frequent reference to the size or weight of objects as determined by yardsticks and scales. Such devices are applied when one wishes to order objects along a quantitative continuum. The question "How much?" is inherent in the process of measurement. Thus, the yardstick not only tells us that one objetct is longer than another; it also tells us *how much* longer it is. In order to do this, the yardstick is divided into equidistant units (e.g., inches) throughout the range of measurement. The measurement of human abilities

1. An earlier version of this bulletin was prepared by Dr. Laurence Siegel, the author's former colleague, now at the Louisana State University.

is quite analogous to the measurement of physical properties. The instrument is a test; its purpose is to order people along some kind of continuum (e.g., subject-matter proficiency); and the units of measurement are read as "scores."

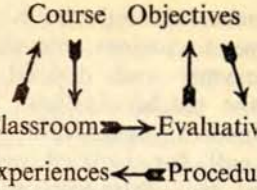
Measurement cannot generally be considered an end in itself. At some point, the scores or readings obtained on the measuring instrument must be interpreted. This process of interpretation or evaluation is implied whenever we speak of objects as being "heavy" or "short' or "cold." It is implied also in an instructor's distinction between students who display a very superior level of performance leading to a grade of A and those who perform at a B level.

Although measurement often facilitates the evaluative process, it is neither the equivalent of, nor a substitute for, evaluation. A final examination may distribute students along a continuum of achievement, but the conversion of test scores to letter grades or even the evaluation of each student's performance with respect to a simple pass-fail dichotomy ultimately rests upon an interpretive process. This process may be subjective; the rationale underlying the evaluative process, however, always involves an element of subjectivity. Thus, it is fairly common to require that students earn a score of sixty percent or better on an examination in order to pass it; this is certainly an objective criterion. The rationale underlying the criterion is nevertheless exeedingly subjective and often indefensible.

It must be recognized, also, that it is possible to evaluate in the absence of formal measurement. We may wish, for example, to evaluate such things as the apparent interest of the student in the subject matter, his attentiveness, and his contributions to classroom discussion, even though we do not yet have tests to measure all such factors.

EVALUATION AND COURSE ORGANIZATION

The evaluative process and the teaching process are both aspects of the same coin. A careful definition of course objectives and the thoughtful organization of classroom experiences are mutually dependent upon each other and upon the preparation and administration of appropriate evaluative instruments. These inter-dependencies have been expressed schematically as follows:



1-6

Course Objectives

Evaluation is a very concrete concept. The teacher does not evaluate in a vacuum; he must evaluate *something*. The things to be evaluated are determined by the objectives set up by the instructor for his course. For the sake of simplicity, let us here assume that a major objective of any course is that of developing competence in the subject area. (Other objectives may be at least as important, but this one will suffice for the purposes of illustration.) The matter of competence is properly evaluated by means of achievement tests.

What, however, do we mean by competence? A run-of the-mill student in Intoductory Psychology may be able to define "retroactive inhibition" without having a real understanding of the implications of the concept. He could trip right over an illustration of it without recognizing it as such! A superior student in the same course, however, has probably seen the relevance of the concept to his own behaviour and the behaviour of others. Undoubtedly, the ability to recognize and apply the concept is the objective toward which the instructor is striving. If this is so, test items dealing with "retroactive inhibition" should be oriented toward application rather than definition. The requirement that a student be able to define the term retroactive inhibition and the requirement that the student be able to show that he can apply the definition represent two different levels of educational objectives. Application is clearly the higher order educational objective. Given the two levels of objectives most instrucotors probably desire that students show proficiency at the higher level (e.g., application). There are, of course, many other objectives held by instructors of an order which are higher than that of application (e.g., analysis, synthesis and extrapolation).

Unfortunately, course objectives in many cases have been narrowly defined. And many objectives which professors hope students will achieve are not systematically included as part of the testing and evaluation plan. The key term is systemati-

cally. While many teachers include the higher order objectives in syllabi, early course discussions, etc. students learn quickly to ignore such statements and depend rather upon their assessment of what the teacher includes in the examination. Thus if a teacher espouses higher level objectives but gives examinations which call for isolated and disconnected sets of facts measured by true false test questions, the student orients his study preparation along these lines and therefore is unlikely to achieve the higher order objectives which the instructor desires.

Classroom Experiences

One of the functions of the definition of course objectives is to provide a rationale for the provision of particular types of classroom experiences. The intent of a course in Physics, for example, to develop an understanding of the significance of objectivity in observation, remains merely a good intention unless students are actually provided with opportunities to observe and record physical phenomena. Similarly, instructors in English are well aware of the fact that the realization of their objective of teaching "effective communication" is dependent upon providing students with an opportunity to write and following this up with a critique of what they have written.

The nature of the learning experiences provided the students will bear rather directly upon the kinds of evaluative techniques judged appropriate for the course. It would be foolhardy to evaluate the laboratory technique of students who have never been exposed to actual work in the laboratory; or to evaluate writing style in a composition course that has concentrated solely upon grammar.

Reorganzing the Course.

It is unlikely that any instructor is ever completely satisfied with a course he teaches. The content of any course is in a constant state of flux; the search for newer and better textbooks is a continual process; the sequence of topics undergoes periodic revision; and new techniques of presentation (including the use of visual aids of various kinds) are incorporated from time to time

The reorganization of course structure is predicted upon two related assumptions: 1) that such reorganization will

make it possible to attain the objectives of the course more satisfactorily; and 2) that the learning experiences currently provided the—student are not as effective as they might be. These assumptions imply that the instructor has evaluated the outcomes of his course and found them lacking. The laboratory portion of the course may not, for example, provide students with an understanding of scientific methology. It may merely be regarded by them as "busy work" contributing little to their knowledge about controls or systematic inquiry.

Thus, we have come around the full circle. Educational objectives and learning experiences both dictate the application of appropriate evaluative procedures and are subject to the basis of the results of such evaluations.

MINIMAL REQUIREMENTS OF MEASURING INSTRUMENTS

Representative Coverage.

It is a fair assumption that *one* of the several objectives of a history course is to convey a basic set of factual knowledge including names, dates, events, etc. The teacher could probably list thousands of bits of information covered in class, the textbook, and outside reading during the year. He could then proceed to write a single test question covering each of these bits of information and administer it to his class in order to determine the extent to which factual knowledge had been acquired by each student. The difficulty in this procedure, however, whould be that the test would contain thousands of questions and would require an inordinate amount of time for construction, administration, and scoring.

In order to overcome these difficulties, the instructor samples each student's knowledge rather than attempting to measure it completely. He may, for example, administer only 150 items covering 150 bits of information. The fact that the teacher willing to generalize from a student's performance on this sample of items to an overall appraisal of the student's knowledge about history implies that there is a substantial correlation between the 150 item test and the exhaustive test covering the full range of historical information. In order for this to be the case, the knowledge sampled by the shorter test must be representative of the full range of knowledge encompassed by the course.

The problem of drawing a sample of test questions from the universe of available questions is paralleled by the problem of sampling for the purpose of public-opinion polling. Suppose, for example, that we wished to determine students' opinions about the semester system. We could if we had unlimited resources and energy, question every student in the school. In order to be practical about it, however, we would decide to question a sample of students. Before generalizing from this sample to the universe (all students) we would need evidence that the sample was, in reality, a miniature representation of the universe. Such factors as age, sex, major field of study, intellectual ability, and grade-average would have to be proportionally distributed in the sample to the same extent that they are found in the total population of all students. Any deviations, or sampling errors, may completely invalidate the results of the survey.

Similarly, errors in drawing a sample of test questions may result in a biased test: i.e., a test that does not effectively measure the full range of knowledge. Such a test is unbalanced. Typically such imbalance exists because proportionately greater weight is assigned to those areas wherein it is easy for the instructor to phrase questions, and less weight is assigned to these areas wherein the instructor experiences difficulty in phrasing items.

The requirement of representative coverage does not, in itself, establish the *number* of questions to be included in a test. Reverting to the earlier illustration of the 150 item sampling of the universe of questions, what would be the effect of maintaining representative coverage but reducing the length of the test of 50 items or 25 items? The optimal length of a test is related to a second requirement of measuring instruments termed "reliability."

Reliability.

A reliable instrument is one that yields consistent readings over a period of time. If we like our roast beef "medium," for example, we would be dissatisfied with a meat thermometer that sometimes caused us to carve the roast when it was rare and at other times when it was cooked to a crisp. The thermometer would be regarded as unreliable because we couldn't depend upon it to give consistent readings.

The concept of reliability is equally applicable in the area of educational measurement. If the scores earned on a

particular test are unstable, the test is unreliable. A perfectly reliable test is one that yields the same ranking of students from best to worst over successive administrations. Thus, if the test were given twice to the same group of students, the highest ranking member of the class on the first administration would also rank highest on the second administration; the lowest ranking student on the first administration would rank lowest the second time; and the intermediate ranking students woud maintain their same relative positions. As a matter of fact, one of the techniques for estimating the reliability of a test is to administer it twice to the same group of persons and to correlate the scores on the two administrations.

In practice reliability estimated by means of the test-retest procedure is only feasible in the case of tests distributed by a commercial publisher. The teacher cannot submit his own tests to this kind of analysis. The requirement of two administrations of the same test wastes precious classroom time and the procedure would undoubtedly be resented by students. Consequently the instructor must content himself with the knowledge that he has developed his test and administered it in accord with certain principles that enhance potential reliability. (There are, however, procedures for determining reliability which are based on only single administration of the test).

A major factor related to test reliability is that of *test length.* A test consisting of just one true-false item would be about as unreliable a measure as could be developed. As the number of items is increased, the reliability of the test increases.

Some clarification of the relationship between the length of a test and its reliability may be derived from consideration of chance as a potential source of unreliability. The uncontrolled factors subsumed under the general classification "chance" lead to correct item responses in the absence of correct information. The contribution of this factor to test scores is most apparent, perhaps, in the case of true-false examinations wherein students who know absolutely nothing about the subject matter being tested are forced to guess for every item. The average score of a group of such students would approximate 50% the possible maximum. It is impossible, however, to predict whether any one of these students will guess correctly or incorrectly at any single item in the test. The only prediction that can be made with a degree

of confidence is that each student will guess correctly about half the time. If we are dealing with a test consisting of just one true-false item, then, we would expect the retest reliability to approximate 0.00 because large numbers of students who guess correctly on the first administration will guess incorrectly on the second and *vice versa*. As the test is progressively lenghtened, we will still obtain this kind of fluctuation for any one item, but the *total* score on the test will become increasingly stable.

This reference to true-false items in no way implies that only objective tests are succeptible to the operation of chance factors. Many uncontrolled factors are operative in the grading of subjective (e.g., essay) eaminations as well. The time of day or night when a particular paper is read, the number of papers that preceded the one presently being graded and the general mood or disposition of the reader may all bear upon his evaluation of a particular student's essay.

Since these factors are not held constant upon retest or rereading, the evaluation of a particular student's response to any one essay question may fluctuate considerably. Again, however, as the test is lengthened by adding additional questions, the total score across all questions will tend to achieve a degree of stability.

It must be recognized that the relationship between test length and reliability does not mean that length is itself an absolute guarantee of reliability. A test consisting entirely of ambiguous or "tricky" items may be quite unreliable in spite of its length. Assuming that a test is well constructed and that appropriate precautions have been taken to insure proper scoring, however, length is probably the most important single factor bearing upon reliability. This fact, if carried to its ridiculous extreme, would cause an instructor to devote all class time to testing. Obviously, practical considerations must enter into decisions about the amount of time to be devoted to measurement.

The foregoing discussion does not necessarily imply that more time must be devoted to testing during the semester. It does, however, have ramifications for the evaluation or interpretation of test scores. A ten-minute weekly quiz, for example, is often useful as a means of motivating students to study. Since it is likely to be an unreliable instrument, however, grade assignment based upon any one of these quizzes should be regarded as extremely tentative. As such

quizzes are administered, the total scores may be cumulated, so that by the end of the fourth week, for example, the instructor assigns grades on the basis of forty minutes of testing and by mid-semester on the basis of eighty minutes of testing, etc. The further into the semester the class goes, the more confident both the instructor and the students can become that the grades based upon this cumulative process are reaching a degree of stability.

Validity.

The validity of a test refers to the extent to which it measures what it is supposed to measure. The fact that a test is reliable, is in itself no guarantee that it will also be valid. A yardstick for example, is a reasonably reliable measuring instrument. It is also quite valid for the purpose of ordering people along a continuum of height. In spite of its reliability however, the yardstick is totally invalid for the purpose of predicting cumulative grade-average. There is no systematic relationship between height and grades and, in consequence no reason to expect a measure of height to be valid against the criterion of grades. This illustration of misapplication of measuring instruments is not as far-fetched as it may appear at first blush. Examinations administered in a course are valid only to the extent that they measure in those areas defined by the course objectives and by the educational experiences provided for the students. This creates some rather obvious difficulties in courses wherein multiple sections are taught by different instructors who give a "departmental" examination. Such an examination is valid only when the instructors concerned are in agreement about the purposes to be served by the course, the topics to be included, the relative importance of each of these topics, and the way in which the presentations are to be made. Such standardization of the teaching process is uncommon and probably undesirable. In consequence, administration of common examinations are indefensible except when such examinations are supplemented by special tests developed by each instructor for administration to his own sections. Unless supplemental tests are administered, the evaluation of student achievements will not reflect the unique flavor given to each section by its instructor.

Equivalence of units.

Earlier in this article educational measurement was likened to physical measurement and parallels were noted between

the achievement test, for example, and the yardstick possesses a characteristic inherent in the term "measurement" but too often neglected in achievement testing. This is the characteristic of equality of scale units. Thus, an object that measures four feet is really twice as long as one that measures two feet in length. The student, however, who scores forty points on an achievement test may know considerably more or less than twice as much as the student who only scores twenty points.

The interpretation (or evaluation) of raw test scores and the conversion of such scores to letter-grade equivalents is a topic deserving attention in its own right. It is sufficient here to indicate merely that as long as we restrict the analysis of test results to a superficial interpretation of the raw scores (e.g., a number of questions answered correctly) we are unable to draw clear distinctions between students, or to order them along a continuum in terms of a unit of measurement indicative of learning and comprehension.