Robert L. Ebel

Measurement and the Teacher

Ten useful principles.

THE principles of measurement of educational achievement presented in this article are based on the experience and research of a great many people who have been working to improve classroom testing. The particular principles discussed here were selected on the basis of their relevance to the questions and problems which arise most often when tests of educational achievement are being considered, prepared and used. While some of the principles may seem open to question, we believe a case can be made in support of each one.

1. *The measurement of educational achievement is essential to effective education.*

Learning is a natural, inevitable result of human living. Some learning would occur even if no special provision were made for it in schools, or no special effort were taken to facilitate it. Yet efficient learning of complex achievements, such as reading, understanding of science, or literary appreciation, requires special motivation, guidance and assistance. Efforts must be directed toward the attainment of specific goals. Students, teachers and others involved in the process of education must know to what degree the goals have been achieved. The measurement of educational achievement can contribute to these activities.

It is occasionally suggested that schools could get along without tests, or indeed that they might even do a better job if testing were prohibited. It is seldom if ever suggested, though, that education can be carried on effectively by teachers and students who have no particular goals in view, or who do not care what or how much is being learned. If tests are outlawed, some other means of assessing educational achievement would have to be used in their place.

2. *An educational test is no more or less than a device for facilitating, extending and refining a teacher's observations of student achievement.*

In spite of the Biblical injunction, most of us find ourselves quite often passing judgments on our fellow men. Is candidate A more deserving of our vote than candidate B? Is C a better physician than D? Is employee E entitled to a raise or a promotion on his merits? Should student F be given a failing mark? Should student L be selected in preference to student M for the leading role in the class play?

Those charged with making such judgments often feel they must do so on the basis of quite inadequate evidence. The characteristics on which the decision should be based may not have been clearly defined. The performances of the various candidates may not have been observed extensively, or under compar-

Robert L. Ebel *is Vice President for General Programs, Educational Testing Service, Princeton, New Jersey.*

able conditions. Instead of recorded data, the judge may have to trust his fallible memory, supplemented with hearsay evidence.

Somewhat similar problems are faced by teachers, as they attempt to assess the achievements of their students. In an effort to solve these problems, tests have been developed. Oral examinations and objective examinations are means for making it easier for the teacher to observe a more extensive sample of student behavior under more carefully controlled conditions.

The price that must be paid for a test's advantages of efficiency and control in the observation of student achievements is some loss in the naturalness of the behavior involved. In tests which attempt to measure the student's typical behavior, especially those aspects of behavior which depend heavily on his interests, attitudes, values or emotional reactions, the artificiality of the test situation may seriously distort the measurements obtained. But this problem is much less serious in tests intended to measure how much the student knows, and what he can do with his knowledge. What is gained in efficiency and precision of measurement usually far outweighs what may be lost due to artificiality of the situation in which the student's behavior is observed.

3. *Every important outcome of education can be measured.*

In order for an outcome of education to be important, it must make a difference. The behavior of a person who has more of a particular outcome must be observably different from that of a person who has less. Perhaps one can imagine some result of education which is so deeply personal that it does not ever affect in any way what he says or does, or how he spends his time. But it is difficult to find any grounds for arguing that such a well concealed achievement is important.

If the achievement does make a difference in what a person can do or does do, then it is measurable. For the most elementary type of measurement requires nothing more than the possibility of making a verifiable observation that person or object X has more of some defined characteristic than person or object Y.

To say that any important educational outcome is measurable is not to say that satisfactory methods of measurement now exist. Certainly it is not to say that every important educational outcome can be measured by means of a paper and pencil test. But it is to reject the claim that some important educational outcomes are too complex or too intangible to be measured. Importance and measurability are logically inseparable.

4. *The most important educational achievement is command of useful knowledge.*

If the importance of an educational outcome may be judged on the basis of what teachers and students spend most of their time doing, it is obvious that acquisition of a command of useful knowledge is a highly important outcome. Or if one asks how the other objectives are to be attained—objectives of self-realization, of human relationship, of economic efficiency, of civic responsibility—it is obvious again that command of useful knowledge is the principal means.

How effectively a person can think about a problem depends largely on how effectively he can command the knowledge that is relevant to the problem. Command of knowledge does not guarantee success, or happiness, or righteousness, but it is difficult to think of any-

thing else a school can attempt to develop which is half as likely to lead to these objectives.

If we give students command of knowledge, if we develop their ability to think, we make them intellectually free and independent. This does not assure us that they will work hard to maintain the status quo, that they will adopt all of our beliefs and accept all of our values. Yet it can make them free men and women in the area in which freedom is most important. We should be wary of an educational program which seeks to change or control student behavior on any other basis than rational self-determination, the basis that command of knowledge provides.

5. *Written tests are well suited to measure the student's command of useful knowledge.*

All knowledge can be expressed in propositions. Propositions are statements that can be judged to be true or false. Scholars, scientists, research workers—all those concerned with adding to our store of knowledge, spend most of their time formulating and verifying propositions.

Implicit in every true-false or multiple-choice test item is a proposition, or several propositions. Essay tests also require a student to demonstrate his command of knowledge.

Some elements of novelty are essential in any question intended to test a student's command of knowledge. He should not be allowed to respond successfully simply on the basis of rote learning or verbal association. He should not be asked a stereotyped question to which a pat answer probably has been committed to memory.

6. *The classroom teacher should prepare most of the tests used to measure educational achievement in the classroom.*

Many published tests are available for classroom use in measuring educational aptitude or achievement in broad areas of knowledge. But there are very few which are specifically appropriate for measuring the achievement of the objectives of a particular unit of work or of a particular period of instruction. Publishers of textbooks sometimes supply booklets of test questions to accompany their texts. These can be useful, although all too often the test questions supplied are of inferior quality—hastily written, unreviewed, untested, and subject to correct response on the basis of rote learning as well as on the basis of understanding.

Even if good ready-made tests were generally available, a case could still be made for teacher-prepared tests; the chief reason being that the process of test development can help the teacher define his objectives. This process can result in tests that are more highly relevant than any external tests are likely to be. It can make the process of measuring educational achievement an integral part of the whole process of instruction, as it should be.

7. *To measure achievement effectively the classroom teacher must be (a) a master of the knowledge or skill to be tested, and (b) a master of the practical arts of testing.*

No courses in educational measurement, no books or articles on the improvement of classroom tests, are likely to enable a poor teacher to make good tests. A teacher's command of the knowledge he is trying to teach, his understanding of common misconceptions regarding this content, his ability to invent novel questions and problems, and his ability to express these clearly and con-

cisely; all these are crucial to his success in test construction. It is unfortunately true that some people who have certificates to teach lack one or more of these prerequisites to good teaching and good testing.

However, there are also some tricks of the trade of test construction. A course in educational measurement, or a book or article on classroom testing can teach these things. Such a course may also serve to shake a teacher's faith—constructively and wholesomely—in some of the popular misconceptions about the processes of testing educational achievement. Among these misconceptions are the belief that only essay tests are useful for measuring the development of a student's higher mental processes; that a test score should indicate what proportion a student does know of what he ought to know; that mistakes in scoring are the main source of error in test scores.

8. *The quality of a classroom test depends on the relevance of the tasks included in it, on the representativeness of its sampling of all aspects of instruction, and on the reliability of the scores it yields.*

If a test question presents a problem like those the student may expect to encounter in his later life outside the classroom, and if the course in which his achievement is being tested did in fact try to teach him how to deal with such problems, then the question is relevant. If the test questions involve, in proportion to their importance, all aspects of achievement the course undertakes to develop, it samples representatively. If the scores students receive on a test agree closely with those they would receive on an independent, equivalent test, then the test yields reliable scores.

Relevance, representativeness and reliability are all matters of degree. Procedures and formulas for calculating estimates of test reliability are well developed, and are described in most books on educational measurement. Estimates of representativeness and relevance are more subjective, less quantitative. Yet this does not mean that relevance and representativeness are any less important than reliability. The more a test has of each the better. While it is possible to have an irrelevant and unrepresentative but highly reliable test, it is seldom necessary and never desirable, to sacrifice any one of the three for the others.

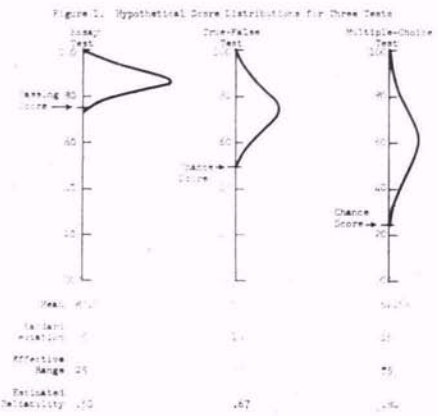
Either essay or objective test forms can be used to present relevant tasks to the examinees. Ordinarily, the greater the novelty of a test question, that is, the smaller the probability that the student has encountered the same question before, or been taught a pat answer to it, the greater its relevance. Because of the greater number of questions involved, it is sometimes easier to include a representative sample of tasks in an objective than in an essay test. For the same reason, and also because of greater uniformity in scoring, objective tests are likely to yield somewhat more reliable scores than are essay tests.

9. *The more variable the scores from a test designed to have a certain maximum possible score, the higher the expected reliability of those scores.*

Reliability is sometimes defined as the proportion of the total variability among the test scores which is not attributable to errors of measurement. The size of the errors of measurement depends on the nature of the test—the kind and the number of items in it. Hence for a particular test, any increase in the total variability of the scores is likely to increase the proportion which is not due to errors of

measurement, and hence to increase the reliability of the test.

Figure 1 shows some hypothetical score distributions for three tests. The essay test consists of 10 questions worth 10 points each, scored by a teacher who regards 75 as a passing score on such a test. The true-false test consists of 100 items, each of which is worth one point if correctly answered, with no subtraction for wrong answers. The multiple-choice test also includes 100 items, each of which offers four alternative answer options. It, too, is scored only for the number of correct answers given, with no "correction for guessing."



Note, in the data at the bottom of Figure 1, the differences among the tests in average score (mean), in variability (standard deviation), in effective range and in estimated reliability. While these are hypothetical data, derived from calculations based on certain assumptions, they are probably reasonably representative of the results most teachers achieve in using tests of these types.

It is possible to obtain scores whose reliability is above .90 using 100 multiple-choice items, but it is not easy to do, and classroom teachers seldom do it in the tests they construct. It is also possible to handle 100-point essay tests and 100-item true-false tests so that their reliability will equal that of a 100-item multiple-choice test. But again, it is not easy to do and classroom teachers seldom succeed in doing it.

10. *The reliability of a test can be increased by increasing the number of questions (or independent points to be scored) and by sharpening the power of individual questions to discriminate between students of high and low achievement.*

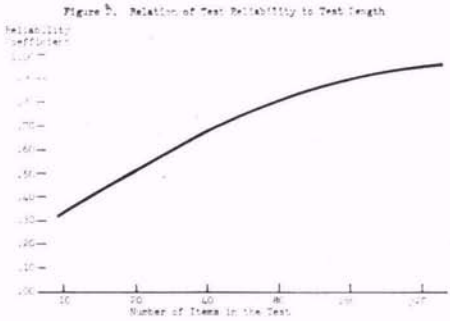


Figure 2 illustrates the increases of test reliability which can be expected as a result of increasing the number of items (or independent points to be scored) in a test. Doubling the length of a 10-item test whose reliability coefficient is .33 increases the reliability to .50. Doubling again brings it up to .67, and so on. These estimates are based on the Spearman-Brown formula for predicting the reliability of a lengthened test. While the formula requires assumptions which may not be justified in all cases, its predictions are usually quite accurate.

Figure 3 shows how the maximum discriminating power of an item is related to its level of difficulty. These discrimi-

(Continued on page 43)

sponsored outside the school system by government and foundation.

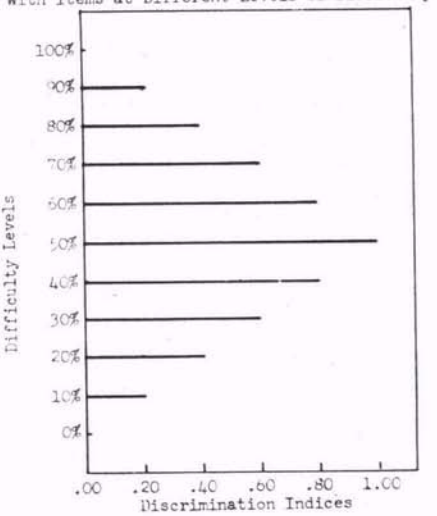
One lead, then, is to redefine the process of curriculum development to make more room for the new functions of both teacher and superintendent. But perhaps the larger question is whether, in acting on our redefinition, we can learn to work in partnership with these colleagues to perform more creatively the functions we may once have thought of essentially as ours—those of keeping purpose sharply in review, of keeping concern for the learner and learning in focus, of relating the selection of content to both purpose and process, and of realizing our best intentions in some kind of balanced perspective. Until we succeed in imagining truly new possibilities for developing capacity and have begun to invent an implementation that may really make a difference, we will continue to be challenged by such proposals and programs as are reported in the publications here under review.

Measurement

(Continued from page 24)

nation indices are simply differences between the proportions of correct response from good and poor students. Good students are those whose total test scores fall among the top 27 percent of the students tested. Poor students are those whose scores make up the bottom 27 percent. An item of 50 percent difficulty does not necessarily have (and usually will not have) an index of discrimination of 1.00. Its discriminating power may be zero, or even negative. But items of middle difficulty have higher ceilings on their discriminating power. What is more important, they not only can have, but usually do have, greater discriminating power than very easy or very difficult items. An item that no one answers correctly, or that everyone answers correctly, cannot discriminate at all. Such an item adds nothing to the reliability of a test.

Figure 3. Maximum Discrimination Attainable With Items at Different Levels of Difficulty



In summary, the 10 principles stated and discussed in this article represent only a sample of the important things classroom teachers need to know about educational measurement. These principles, and the brief discussion of each presented here, may serve to call into question some common practices in classroom testing, or to suggest some ways in which classroom tests might be improved. They are not likely, and are not intended, to say all that needs to be said or do all that needs to be done to improve educational measurement in the classroom. It is our sincere belief, however, that a teacher whose classroom testing reflects an understanding of these principles will do a better than average job of measuring student achievement.