Research in Review

Column Editor: Ira J. Gordon

Testing and Evaluation: Pupils and Programs

I HAD just turned in my summer session grades. As usual, a portion of the grade was based upon a "teacher-made" test. In the last class period, an item analysis had been performed by show of hands. The top and bottom fourth of the class indicated their success on each item, and those questions which showed that a statistically significant higher proportion of the top group were correct were considered valid or good. Why? We had learned that a test item, to be good, must discriminate.

We ran into problems, and these constitute the basis for this discussion. Not many items survived the analysis. This led to questions such as these: Is the purpose of testing only to discriminate "good" from "poor" students? Why is an item that all answer correctly a poor test item? Even if a test discriminates, does it really measure what was learned? Are testing and evaluation synonymous? What assurances are there that a test measures the desired educational outcomes?

In the field of programmed learning, any teaching item or question which a student misses is a bad item. The aim is for near perfect success. When testing occurs, if all students answer all items correctly, and these items measure the content of the program, then the program is a good one, and the test a good measure. In such a case, the aim of measurement is the effectiveness of a program, not the assignment of grades to pupils. According to programmers, failure in learning is avoided by successful programming.

If the classroom teacher were to use this criterion, any test item which measured what was taught would be acceptable. The crucial question would be whether the test measures what was taught, rather than whether it discriminates between high and low students.

Measure of Change

Let us extend this to the evaluation of new curricular developments, rather than to the classroom teacher's homemade test. How can new programs be evaluated? As an example, the newer approaches to mathematics stress discovery techniques and the use of concrete materials. Can such a program be measured by standard arithmetic achievement tests, designed for different educational practices? The new science programs face the same basic problem.

Evaluation consists then of a valid measure of change or learning brought about by the educational procedure. Curricula using "discovery" techniques can-

not be evaluated by tests which do not include discovery items, nor can pupils who have not been taught such techniques be measured by discovery items. The new curricula stress procedures, approaches to learning, learning to learn. How well can these goals be measured? Simple discrimination on test results between those who were in the "new" program and those in the "traditional" program may conceal the value of either approach.

The simple experimental design of one control group and one experimental group may therefore be inadequate. We may develop test items which discriminate, simply because the programs differ, but which do not truly demonstrate the superiority of a program.

There is a need to engage in a more rigorous, less "bandwagony" look at the evaluation of new programs. We have long known of the Hawthorne effect, in which people do better simply because they are part of an experiment. How often are results published which indicate that any change brings about increased learning? The failure to engage in rigorous control over the morale effect has often led to fallacious results.

The problem of evaluation is also complicated by the confusion between innovation and experimentation. We often term the former the latter. We try something new and like it, and then assume we have conducted an experiment which demonstrates that our innovation really works. Rigorous evaluation, with careful controls, clear hypotheses, and a clear understanding of the many variables involved in any educational experimentation may move us toward better curricular research.

What is of concern is the assessment of change as a result of a planned learning experience. In order to do so effec-

tively, each of the following aspects of the problem must be considered: "(a) obtaining appropriate measuring instruments, (b) securing consistently good rapport with examinees, (c) selecting tests with appropriate norms, (d) estimating the amount of change and its statistical significance" (Davis, 1962, 11).

Further, new programs are often tried out on special segments of the school population—the gifted, the culturally deprived, etc. Testing and evaluation of these programs is often naive, overlooking various sampling errors. As Lord indicates, some of the problems in assessment of growth include comparing gains made by students who obtained low initial scores with gains made by students who obtained high initial scores and comparing numerically equal gains made by students with different initial scores (Lord, 1958). Our IBM computers allow designs which enable us to surmount this type of problem easily, *if we design our research correctly.*

Both the classroom teacher and the curriculum researcher, therefore, face common problems. They both have to ask what am I measuring *for*? What are my goals? Are my tests adequate to measure the achievement of these goals? What should be my criteria for the validity of my tests? Am I concerned with individual growth and grading, or the evaluation of the effectiveness of a planned learning program?

Another aspect of testing and evaluation concerns those variables which should be measured, yet often are not. If an educational program is supposed to yield changes in social and emotional growth and in heightened motivation, is it legitimate to measure only growth in academic achievement? Admittedly, personality variables are hard to handle, especially in school situations. Yet, they

are often listed as important outcomes, without careful measurement.

The work on creativity (Torrance, 1962) offers leads into the problem of the measurement of formerly vague and ambiguous terms. The first step is still definition. In the case of careful research, these definitions must be operational. They must be capable of being used by various researchers in reliable ways. Clear-cut evaluation requires operational definitions of personality variables such as motivation, self-concept, interest, social adjustment, mental health. When this has been accomplished we will be able to shift from making either vague claims for the success of programs or the virtual neglect of their measurement. We will probably find these terms too huge, or useful only as broad constructs, and will substitute sub-variables for them. This has already occurred with the term "mental health." We will also explore the relationships among these variables, academic achievement, and the teaching-learning process.

We may end up with our old slogan: each child learns in his own fashion, and no single approach is best. If we do, it will be with an added power: we will have developed the tools to assess what programs work for which children, and why.

References

Frederick B. Davis. "Testing and the Use of Test Results." *Review of Educational Research* 32:5-14; February 1962.

Frederick M. Lord. "Further Problems in the Measurement of Growth." *Educational and Psychological Measurement* 18:437-51; Autumn 1958.

E. Paul Torrance. *Guiding Creative Talent*. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1962.

—Ira J. Gordon, *Professor of Education, University of Florida, Gainesville, Florida.*

(Continued from page 37)

activity. Here, effort and accomplishment are quite apparent. The "feel" of the muscles when a movement is performed, the appearance of the body (even to the extent of looking in the mirror) that accompanies the control of those muscles, all tend to give the student an instantaneous recognition of effort and accomplishment. What is more, the teacher, too, is aware of his achievements and evaluations at the same time. The possibilities of introducing advanced work prematurely are less likely in such circumstances than in those situations in which communication of results is not as readily apparent.

It seems to me that it is important for educators to note clearly the difference between self-evaluation and grading. The former is a learning process, the latter a competitive placement. I hold no brief against competition. In its place, it is a great stimulus for one to work harder, to increase even further his performance. I would like to make clear that a competitive drive without the ability to evaluate one's own accomplishments may become a frustrating experience. The "naturals" sense all this themselves and proceed to greater and greater heights. It is the larger proportion of eager and conscientious individuals, geared to a routined set of directions, who must be taught how to move or think creatively, and then how to evaluate their learning. In essence, it is a consciousness, rather than a "self-consciousness," of one's self. I maintain that self-evaluation is not an art or a skill inherent in all of us, but is a method of study which must be taught in all the early stages of education.