ONGOING EVALUATION OF EDUCATIONAL PROGRAMS*

Stephen P. Klein

CSE Report No. 83

October 1972

Evaluation Technologies Program
Center for the Study of Evaluation
University of California
Los Angeles, California

*Prepared for presentation at a symposium on "The Relative Strengths of Norm-Referenced and Criterion-Referenced Achievement Tests," 1972 Convention of the American Psychological Association, Honolulu, Hawaii, September 5, 1972. The author's formulation of many ideas expressed in this paper was facilitated greatly by his discussions of the topic area with Professor Eva Baker and Professor W. James Popham of UCLA.

When I was invited to participate in this symposium, I was asked to talk about the nature of criterion- and norm-referenced measurement with respect to the ongoing evaluation of educational programs. This invitation may have been stimulated by a desire to balance the symposium in that most of the other participants have had a long association with norm-referenced measurement, while my affiliation with UCLA and my assigned topic would almost ensure my being a strong proponent for the criterion side of the controversy. Actually, this is not the case, since I have refrained from involving myself in this issue.

The reason for my abstinence is that the arguments on both sides of the controversy are irrelevant when they are compared to what we know about educational measurement in the real world. As a case in point, Husek and Popham (1969) have stated that "Norm-referenced measures are . . . used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device . . . Criterion-referenced measures . . . are used to ascertain an individual's status with respect to some criterion, i.e., performance standard." Unfortunately, this frequently cited distinction between norm- and criterion-referenced measurement does not make a great deal of sense, since the same measuring device can be used for both purposes (Klein, 1970). That is, one can use the same scores on a given measure both to compare the performance of different students and to assess whether each student has achieved a certain level of proficiency. Professional competency measures such as the Real Estate and Bar examinations illustrate this double use. Scores on these kinds of tests are

often reported in terms of each student's rank as well as whether a given score constitutes passing.

A second supposed distinction between these two kinds of measures is that only criterion-referenced instruments deal with specific objectives. This is hardly the case. My colleagues at UCLA (Skager, 1971) are in the process of building measures at various levels in a hierarchy of objectives by obtaining a representative sample of test items from measures used to assess performance at lower levels of this hierarchy. Measures constructed at all levels are based on well defined objectives; it is just that some of these objectives are more general than others. Thus, if one of the supposed differences between norm- and criterion-referenced measures is that norm-referenced tests are more general in nature, then it appears that this distinction involves a purely arbitrary judgment based on test length and the number of subobjectives assessed. There is no fundamental difference in the nature of the measures themselves. Some people think there is, however, simply because short measures of only a few specific objectives are used for assessing relatively small units of instruction, while the more general measures are supposed to assess student performance across a much wider range of skills and knowledge (Klein, 1971). Both kinds of measures are still based on objectives.

A third factor that is supposed to differentiate norm- and criterion-referenced measures is the set of decision rules used in retaining or deleting items from the final form of the measure. Both kinds of measures should contain only items measuring the objectives

covered by the test, but the norm-referenced instrument is also supposed to have items that differentiate among students of varying levels of ability or knowledge. Items that are too easy or too hard are deleted. An item's difficulty, of course, is not something inherent in the item itself but, rather, is a function of the student's knowledge, ability, and prior experience at the time he encounters that item. Thus, the item "2 x 4 = ?" may be difficult for some students at one point in their education, but very easy for them at subsequent points. If a student gets this item correct, it is also likely that he will get the item "2 x 2 = ?" correct, but we would have to know more about the student's age and education to predict well whether he would answer correctly the item "2 x 7 = ?" even though all of these items deal with the same objective, namely: the multiplication of two single-digit positive integers. If a student answered all of these items correctly or failed all of them, we would have gained relatively little information about his performance capabilities. If, on the other hand, we were able to find that point on the scale of item difficulties for a given objective at which he answered all the items correctly but beyond which he could not perform in terms of getting more difficult items correct, then we would know at what level the student had mastered the objective. The trick, therefore, is not to construct a test to just measure an objective but, rather, to construct one that pinpoints most accurately a given student's performance level with respect to that objective. Thus, the wider the scope of the abilities and experiences of students taking a test, the wider the range of item

difficulties that have to be included to ensure a full scale has been developed. Using several test items of the same difficulty for a given objective is, therefore, a rather uneconomical use of testing time (unless, of course, the items are very unreliable). This is true of both the so-called norm- and criterion-referenced measures and points out that there should be no difference in their modes of construction. In those rare situations in which information is desired only about whether students have reached a certain performance level on a given objective, then one need only construct two or three items at that level to assess whether it has been reached. It would not be necessary to have a separate measure for this purpose.

The foregoing discussion has indicated that there is no fundamental difference between norm- or criterion-referenced measures in terms of their focus, foundation in objectives, or in their desired modes of construction. The difference between norm- and criterion-referenced approaches, therefore, is not in the measurement, but in the interpretation of the results of that measurement. One can now ask whether ongoing evaluations should emphasize one or both of these two kinds of interpretations. I contend that both are required. This can be seen by examining four quite diverse purposes for which an ongoing evaluation might be conducted. These purposes are as follows:

1. Identifying program components that need improvement.
2. Identifying students needing special attention.
3. Providing the basis for teacher accountability and school accreditation systems.
4. Determining whether a program is being implemented as planned.

Identifying Program Components Needing Improvement

Program and curriculum developers are among the most avid proponents of the criterion-referenced approach, especially when it comes to identifying program components needing improvement. They contend that one must first state the objectives an instructional program should strive to achieve and then measure student performance with respect to these objectives. If a given component of a program fails to yield the desired changes in student performance, one must then modify this component until it works successfully.

I followed this logic in developing a training workshop in evaluation (Klein, et al., 1971; Klein & Nadeau, 1971). This process involved constructing an initial draft of the workshop, field testing it, making revisions, field testing it again, and so on until I was satisfied that most of the people who participated in the workshop would do well on each of its objectives. Although this sounds like a strictly criterion-referenced approach, I frequently relied on normative data to help make decisions about the program. This often took the form of a question about whether a given change in the materials, entailing much longer instructional time, was worth the increase in scores that would accrue from it. Similarly, if two alternative ways of presenting certain information took the same amount of instructional time, but one yielded higher test scores, I would have had a good basis for deciding which one to use. In short, comparative data about alternative program units as well as criterion-referenced data about a given unit's success was useful and necessary for the purposes of product development.

Identifying Students Needing Attention

A second purpose for conducting ongoing evaluations is to identify students or groups of students needing special attention in the sense that they are not achieving some or all of a program's objectives. It is interesting to note, in this context, that if all the students in a program were not achieving a given objective, then this purpose would not be an issue. In other words, one would not look at an individual student but, instead, would begin to question the efficacy of the whole program. Now if we follow this logic one step further, it becomes obvious that we are making norm-referenced interpretations in the sense that we are comparing what is happening to some students versus the reference group of all the other students.

This situation was illustrated well during the development of the Southwest Regional Laboratory's (SWRL) new reading program. For those who may not be familiar with this program, it has one facet of particular relevance to the so-called issue of norm- versus criterion-referenced measurement. This aspect involves a criterion-referenced pacing, monitoring, and reporting system that actually improves the rate and level of student achievement in the program (Sullivan & Niedermeyer, 1972). A salient feature of this system is a matrix of students by objectives in which the cell entries indicate the degree to which each student is achieving each of the program's objectives over the period of the program's operation. One need only glance at these charts to identify students and groups of students who are mastering the program's objectives relatively slowly or quickly.

When a SWRL staff member was field testing this reading program in one school district, he noted that all but one of the kindergarten classes were on schedule in meeting the program's objectives. When the teacher whose class had the discrepant results was confronted with this information, she claimed she had not tried to implement the program because she felt it was too hard for her students. The SWRL man, despite his training and experience in the bastions of the criterion-referenced school of thought, replied: "Your students are no less able than students in other kindergarten classes in this school; if the students in these classes can do the work, why can't your students do it?" (Niedermeyer, 1972) The only normative data left out of this argument was the transformation of the average rate of student mastery of the program's objectives into grade-level equivalents.

This case points out a common misconception with respect to criterion-referenced measurement; namely: the supposed lack of variance in student scores. This misconception has been brought about by the failure to recognize that student performance is a function of both the level and rate of student learning. If the focus is on level, then the scores should be reported in terms of number or percent correct since rate has either been held constant or ignored (such as testing all students on a given date). If the focus is on rate, then scores should be reported in terms of the time it takes to achieve a given criterion standard, i.e., since level is held constant. The following two situations involving selection decisions illustrate these differences in foci as well as the interrelationships of level and rate:

a. Applicants for a typing job are selected solely on the basis of their relative typing proficiency. The time it took an applicant to gain the level of proficiency needed to be hired is ignored.

b. A program to train pilots retains only those students who have mastered a given criterion level in the first 10 hours of instruction. This is done because experience has indicated that students who fail to progress at this rate will take too long to complete the entire training program.

The foregoing kinds of situations have led to assessing variance among students in the norm-referenced approach by means of computing differences in the levels of performance, while in the criterion-referenced approach the variance among students is assessed by computing differences in the times it takes to achieve various performance levels. Thus, there is variance in student performance in both approaches; the question is to know where to look for it. If rate is held constant, the variance is in level; if level is held constant, the variance is in rate.

The public's desire for having student performance reported in terms of grade-level equivalents (despite the many shortcomings of this type of scale) may result from the fact that grade equivalents appear to combine both level and rate of performance. Indices such as percentiles, standard scores, and the number of objectives achieved by various percentages of students may not have received such acceptance because they do not integrate the concepts of level and rate.

Accountability and Accreditation Systems

California, along with many other states, is beginning to institute various laws requiring teachers to be held accountable for their students'

performance in meeting various educational objectives. This trend has led to a third purpose for ongoing evaluations of educational programs; namely: providing the data base for making decisions about the adequacy of the performance of the students in each teacher's class. A corresponding set of practices for accrediting schools is being considered by several other state legislatures and departments of education. All of these systems consistently acknowledge two important factors that must be included in any accountability or accreditation approach that relies on test data; first, student performance should be measured and the scores reported with respect to specific goals or objectives, and second, there must be some control or adjustment for initial differences in student ability and performance (Popham, 1971; Klein, 1972; Klein & Alkin, 1972). These two considerations are, of course, respectively the essence of the criterion- and norm-referenced approaches. Further, reports to the public regarding the results of such systems will contain both types of information. In the case of accreditation, for example, the results would be reported in terms of whether a school's students were performing at an acceptable level, i.e., criterion standard, as well as in the normative terms of whether the obtained student performance was below, at, or above a level that could reasonably be expected of them, given their abilities, skills, and attitudes prior to entering the school.

Implementation Evaluation

Earlier in this paper I noted that one of the purposes of conducting an ongoing evaluation was to obtain data on each of a program's objectives

so as to identify the particular program procedures that may be causing differential performance levels among students. This is done to provide information for correcting deficiencies as well as to obtain wider use of the procedures that have been shown to be more effective. An analysis of a program's procedures, on the other hand, might indicate the various kinds of processes that deserve attention in terms of their potential for producing differences in student performance. An analysis of these procedures would involve the use of observations, checklists, unobtrusive indicators, and related measurement techniques. The results of such an evaluation would, of course, be reported in both norm- and criterion-referenced terms. This would entail describing whether a given procedure was or was not implemented along with an indication of the degree or quality of this implementation.

One study that illustrates the importance of using norm- and criterion-referenced interpretations of implementation data involved evaluation of a school district's teacher appraisal and improvement plan (Niedermeyer & Klein, 1972). This plan was being tried out in five of a district's 26 elementary schools. One of the important goals of this plan was that it should encourage teachers to focus their attention on student performance rather than just on the instructional techniques and materials they employ. This corresponds to the distinction between product and process criteria. An investigation of whether or not this goal was achieved was carried out in part by asking a sample of all the teachers in the district the following question:

> Briefly describe two examples or indicants that illustrate how your teaching effectiveness improved this year or has been improved over previous years due to the teacher evaluation system in your school.

A content analysis of the answers was done to determine the extent to which a teacher replied in terms of processes employed versus student performances obtained. The results of this analysis indicated that for teachers who were trying out the new program, 50% emphasized a product orientation and 50% were neutral or emphasized the process orientation. In order to determine the significance of this result, however, one must compare these data to that obtained with the teachers who were not using the new system.* In short, one often needs a normative frame of reference in order to help set reasonable criterion levels.

Summary and Conclusions

In this discussion I cited four quite diverse purposes for conducting ongoing evaluations of educational programs. These purposes varied considerably in their scope and emphases, ranging from yearly periodic reporting of summary test results to almost daily recording of individual student performance, and from an emphasis on outcomes and product development to an emphasis on process variables. Despite this diversity, there is clear evidence that one needs both norm- and criterion-referenced interpretations of the data obtained in such evaluations in order to make realistic appraisals of the program being investigated.

---

*The results for teachers in the traditional evaluation system were 13% and 87%, respectively.

REFERENCES

Husek, T. & Popham, W.J. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6 1-9.

Klein, S.P. Evaluating tests in terms of the information they provide. Evaluation Comment, 1970, 2 (2), 1-6.

Klein, S.P. The uses and limitations of standardized tests in meeting the demands for accountability. Evaluation Comment, 1971, 2 (4), 1-7.

Klein, S.P., Burry, J., Churchman, D.A., & Nadeau, M. Evaluation workshop I: An orientation. Monterey, California: CTB/McGraw-Hill, 1971.

Klein, S.P. & Nadeau, M. The development and field testing of Evaluation workshop I: An orientation. CSE Report No. 71. Center for the Study of Evaluation, UCLA, September, 1971.

Klein, S.P. & Alkin, M.C. Evaluating teachers for outcome accountability. Evaluation Comment, 1972, 3 (3), 5-11.

Klein, S.P. An evaluation of New Mexico's educational priorities. Paper presented at the 1972 Convention of the Western Psychological Association, Portland, Oregon, April 4, 1972.

Niedermeyer, F. & Klein, S.P. An empirical evaluation of a district's teacher accountability system. Phi Delta Kappan, 1972 (in press).

Niedermeyer, F.C. Personal communication, July, 1972.

Popham, W.J. Designing teacher evaluation systems. Los Angeles: Instructional Objectives Exchange, 1971.

Skager, R.W. The system for objectives-based evaluation--reading. Evaluation Comment, 1971, 3 (1), 6-11.

Sullivan, H.J. & Niedermeyer, F.C. Pupil achievement in reading under varying levels of teacher accountability. Submitted to AERJ, June, 1972.