

LOCAL EVALUATION AND SCHOOL IMPROVEMENT:  
CURRENT STATUS AND FUTURE POSSIBILITIES

Joan Herman

CSE Report No. 246  
1985

CENTER FOR THE STUDY OF EVALUATION  
Graduate School of Education  
University of California, Los Angeles

The project presented or reported herein was supported pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

## Table of Contents

	<u>Page</u>
Introduction	1
Testing in the Schools	2
The Benefits of Required Testing	3
Reasons for Disuse	4
A Basic Validity Problem	6
References	14

## Introduction

Evaluation blossomed in the mid-sixties with the bright promise of bringing order, accountability, and improvement to the world of public education. It would provide, we thought, a powerful methodology for solving a variety of problems. We would define our goals; we would assess needs relative to those goals; we would plan programs to address the needs and goals; we would implement our well targetted programs, assess their progress and outcomes, and learn from experience (and our data) about how to plan and implement better the next time, discarding programs and ideas which didn't work and refining those which did. It was an appealingly simple idea and a tribute to the presumed rationality of the educational system.

Two decades later, evaluation withers. It has spawned considerable interest in achievement testing -- witness the growth of standardized testing, of minimum competency testing, of district continuum testing, of curriculum embedded testing, of state and national assessments--but its effectiveness is suspect. Lots of testing goes on in the names of evaluation and scientific management, but to what extent do the results get used and to what extent do they stimulate the improvement of education? Research documents some modest effects on program-level decision-making (Alkin et al, 1979; Cohen & Garet, 1975; King & Pechman, 1982; Weiss, 1972; Worthen & Sanders, 1973), but evaluation has yet to reach its promise in improving schools.

This paper explores various issues in the use of evaluation data by local schools. Based on a recently completed national survey of teachers and principals, the paper first documents how much achievement testing typically goes on in local schools, its costs and benefits. Next,

potential reasons impeding the usefulness of tests are discussed. Finally, alternative models are presented for increasing the efficacy of evaluation and its role in school improvement.

### Testing in the Schools

Although those in the profession might disagree, evaluation in schools has been synonymous with achievement testing. And evaluation, by this definition, has undergone rapid growth in the last decade. A few figures attest to this growth: according to the General Accounting Office, 90 percent of the local educational agencies throughout the country administer standardized, norm-referenced tests to students within their purview (HRD-76-113), supplemented in many cases with district curriculum testing; 42 states conduct state assessment programs (Hall, 1975); 38 in addition have adopted minimum competency testing requirements (Gorth & Perkins, 1979) and various special state and federal programs have added additional mandates. This results in considerable required testing in the schools.

How much? A recently completed study conducted by the UCLA Center for the Study of Evaluation with funding from the National Institute of Education provides some basic answers. (See Dorr-Bremme, Herman, & Doherty, 1983.) Based on a survey of principals and teachers in a representative sample of 100 districts throughout the country, the study found that the typical upper elementary school student spends about six hours a year taking required tests in reading and about five and three-quarters hours taking required math tests. Figures at the secondary school level were slightly higher, with the typical student spending about seven hours a year in mandated English tests and six hours in required math tests.

These figures, it should be noted, reflect a very conservative estimate of the amount of time devoted to testing. They only include teacher estimates of test administration time, excluding the substantial preparation and wind-down time frequently associated with formal testing occasions. Dorr-Bremme et al (1983) found that actual administration time accounted for only 60% of student time devoted to testing, and that for every one hour of student test time, teachers typically spent two to three hours in test related activity, e.g., preparation, scoring, recording, etc. Testing, then, can consume considerable resources in teacher and student time, time that may represent important opportunity costs for instructional programs.

#### The Benefits of Required Testing

The costs of district evaluation and testing programs must also be considered in relationship to their benefits. What are the benefits for school and classroom practice? The results of CSE's national study again provides some answers..(See Dorr-Bremme, 1983; Herman & Dorr-Bremme, 1983.) Consonant with the findings of other researchers ( Airasian, 1979; Goslin, 1965; Goslin et al, 1965; Salmon-Cox, 1981), CSE results indicate that teachers and principals generally pay relatively little attention to required tests. While these tests do provide some general information for principals on the strengths and weaknesses of their school's curriculum and do serve as a general benchmark against which teachers can assess students needs and progress, they have little direct impact on instructional and school decisionmaking --and this appears equally true for district continuum or objectives based tests as well as standardized, norm referenced test batteries. An exception is in schools serving students of

low SES. Here test content seems to be driving the curriculum and principals and teachers appear under substantial pressure to produce results. More generally, however, test results play a lesser role. From the principals' perspective, for example, test results play a minor role in allocating financial resources to areas of need, and contrary to teacher fears, reportedly contribute little, if anything, to teacher evaluation. From the teachers' perspective, the results of required tests figure very little in routine classroom decision-making. Not surprisingly, it is their own teacher developed tests and, at the elementary school level, those which accompany curricular materials which figure most heavily in instructional planning.

#### Reasons for Disuse

Why does this formal evaluative data remain untapped by schools? The reasons are many. First, most formal testing programs are designed without adequate attention to the qualities which are important to educators. Herman and Dorr-Bremme (op cit) have identified several features which increase the usefulness of tests for teachers, features which characterize the assessment instruments which teachers use most frequently:

Validity: tests must assess what the teacher believes he or she has actually taught in a way that seems consonant with the way he or she has taught it;

Suitability: their intended purpose must fit the tasks the teacher needs to accomplish; they must be flexible to meet student, teacher and classroom needs;

Immediate availability: the teacher must be able to employ the tests whenever it seems appropriate to do so and have the results back promptly; and

Ownership and control: related to the previous three factors, the teacher must believe that that the test represents his or her needs and aims and that it is a useful tool.

In short, the assessment tools that teachers need must be sensitive to local conditions and to the particular circumstances in their classrooms. Principals too need assessment tools of the same type in order to serve as instructional leaders in their schools. It is not surprising then that both teachers and principals rely heavily on assessment strategies that are internal to the school and its classrooms: teacher-made tests and assignments, teachers' observations and judgments, and the readily available, instructionally relevant tests that come with commercial curriculum materials. These classroom-based instruments exemplify the essential features summarized above. In contrast, externally mandated measures generally do not. The fit between their content and format and a particular teacher's curriculum often is moot. Test scores often are not returned until months after administration. Too often, too, the results come back in a format teachers and many principals find difficult and confusing. For these reasons, the results of standardized tests, other minimum competency measures, and many district continuum tests can seem remote and irrelevant to teachers and principals.

Educators' background and learning in the area of testing and evaluation also may help to account for their use of formal measures. A variety of studies indicate that teachers and principals have very limited formal training in the areas of evaluation, testing, or measurement (Yeh 1978; Rudman et al, 1980) While most districts provide directions on how to administer district tests and do provide some reports --albeit sometimes

quite belatedly -- of test results, few teachers receive training or assistance in how to use tests to improve their instruction or in other evaluation methods (see Herman & Dorr-Bremme, 1983).

Related to knowledge and background are teachers' attitudes toward required tests. The CSE study indicates that lack of familiarity does breed some contempt, i.e., where more staff development and training in testing was observed, teachers were more positive about the quality and utility of tests, and test results had greater influence on decisionmaking (Dorr-Bremme, Herman, & Doherty, 1983). Further, while most teachers and principals who participated in CSE's survey were relatively positive or neutral toward testing, a sizeable minority were concerned about the quality of available commercial and district tests, their match to the school and their classroom curriculum, and their fairness for some students.

These concerns are mirrored by some in the research community. A number of research studies, for example, have been discouraged by the quality of commercially available tests, both norm-referenced and criterion-referenced varieties (CSE, 1974, 1976, 1979; Huron Institute, 1978).

#### A Basic Validity Problem

Content validity is certainly a chief quality issue. Tests are valid, i.e., accurate and appropriate, indicators of students and school performance only under special circumstances. The test needs to reflect the actual program of study children receive. If an examination tests content and skills at odds with school experiences, then the information provided is very weak, and a fatal break occurs in the logic and validity

of any evaluation. Evaluation assumes a rational system, i.e., that the school systems define goals and objectives, that their instructional materials and their teachers in fact teach these goals and objectives, and finally that the system is assessed with tests which actually measure these same goals and objectives. Unfortunately, reality does not match this ideal, as a case study from a local California district demonstrates (See Herman, 1984). This school district, like many others around the country, was under tremendous pressure to raise its test scores. Students in the district annually took both the California State Assessment (known as CAP) and a standardized test. Students consistently scored well below average --and below expectation--on both measures. The school board was disgusted and the assistant superintendent was desperate. We suggested that the district examine whether they were teaching the skills that were being assessed on CAP and develop instruction and practice exercises where any gaps existed. The analysis of the match between the third grade test in reading and the third grade basic reading text (which, as in many districts, de facto defined the district curriculum objectives), produced a startling discovery: the third grade text provided direct instruction and practice in less than one-half of the skills included on the test, placing students and the district at a distinct disadvantage. (We provided supplemental practice for students in these areas of neglect and made some other suggestions regarding the testing procedures; while causality is difficult to attribute, the results were positive.)

Later, when this same school district was ready to select a new standardized test, we suggested that they would be well to pay attention to curricular match. That is, they should choose a test whose objectives

matched those of the district curriculum. Since they continued to be concerned about their performance on the California Assessment Program, they might also want to optimize the match between the two tests. In other words, if they wanted their students to do well on both tests, they should be sure that their curriculum, to the extent possible, prepared students in the tested skills. We therefore did an analysis of the match in reading, math, and language arts between the district curriculum objectives, the California Assessment objectives, and three commonly used standardized tests. Again, the results demonstrated the irrationality of our current educational system (See Herman & Cabello, 1983). First, in some cases a standardized test covered less than 20% of the district's grade level curricular objectives and 20-30% coverage occurred for half the tests (a small percentage for judging the quality of the district's program). Second, from 40-100% of the items found on the standardized tests were in fact covered by the district curriculum. A sizeable proportion were around the 60-75% range, leaving 25-40% of the test at odds with the curriculum and students' instructional experiences. Finally, given the match between the curriculum and the California Assessment Program, it probably will come as no surprise that there was a generally poor match between the objectives on CAP and those on the standardized tests, from a low of 17%, admittedly to a high of 100%. The point is, under these conditions, it is difficult for students to perform well and the problem is exacerbated by the fact that tests which provide a good curricular match at one grade level or in one subject are not necessarily the ones which are best at other levels and subjects. Is it any wonder that educators often feel that formal testing is irrelevant to their needs and that these evaluations have little meaning for them?

Evaluation, under these circumstances, seems to be something that is done to, rather than for or with the schools--and many schools are in fact "done in" by the effort. Baker (1983) has discussed extensively the fundamental problem inherent in the evaluation situation: the mismatch between the evaluation needs of local schools and those of policy and accountability interests. Formal evaluation measures are driven by external demands irrelevant and insensitive to local perspectives and problems, and their results frequently are ignored as invalid at the local level. Meanwhile, teachers and principals persist in their routine instructional and assessment practices, which are ignored as unreliable by policy levels. The result, according to Baker, is an inefficient layering of overlapping evaluation requirements: superordinate (federal and state) demands; regular and special district demands; school-imposed requirements and classroom driven needs.

But the situation is not totally bleak, as research conducted by Bank and Williams (1981) demonstrates. Their work indicates that influential idea champions (key administrators who are committed to improving instruction through evaluation) working with a stable core of staff (representing a critical mass of teachers and administrators) and oriented toward comprehensive problem analysis can make a difference in evaluation practices and instructional effectiveness.

Bank and Williams (1983) describe an example from one district they studied. This district, with the help of its teachers over a period of years, developed both a district curriculum continuum and a pool of test items keyed to the curriculum at each grade level. Pre- and post-tests were then designed to monitor school and student progress and to help

teachers provide more effective instruction targeted at areas of student need.

The district, however, recognized that a testing program by itself could not improve instructional practices. The central office therefore instituted a staff program to train teachers in a diagnostic-prescriptive approach to teaching and also created a new school-based role, resource teacher, to help teachers implement the system and provide more individualized instruction. The results of this effort over a ten year period (Bank & Williams, 1983) support district claims that the testing system holds the curriculum, instructional materials, staff development, and teacher supervision together. The system works as follows:

- All tests are directly linked to a K-8 instructional continuum for reading, language arts, and mathematics. They are the primary tools for monitoring instruction.

- The tests cover a full range of skills, from minimum to maximum competency;

- The tests are administered quarterly, or more often at a teacher's request.

- The tests are computer-scored. Results are formatted in easy-to-read, specially tailored reports to teachers, parents, and site and district administrators. Teacher printouts, for example, are returned with their students scores organized by objectives and printed out by learning group.

- The teacher uses year end results to set instructional priorities and objectives for individual students and for learning groups; progress is

monitored periodically by both the principal and the teacher; and appropriate actions are taken immediately to remediate children who are not performing well;

- The school principal is expected to monitor the progress of students and classrooms. The principal is also expected to spend time observing teachers in their classrooms. During end of year planning days, the principal meets with the teachers to discuss the expected progress of groups of students for the year. During the course of the year, principals receive student scores and assess progress. Teachers are held accountable for student performance; not all students must achieve the preset goals, but deviations must be explained.

- Principals meet regularly with the superintendent to discuss student progress.

- The tests are reviewed and revised on an on-going basis.

The result is a multipurpose, integrated system which helps teachers provide better, more individualized instruction and one which facilitates accountability and serves the needs of policy makers as well. It exemplifies some important features for increasing the utility of tests and evaluation systems for teachers:

- the tests match the curriculum as teachers perceive it;
- the tests were developed by and for teachers, and teachers feel considerable ownership of the system;
- the timing of the tests is under teacher control (beginning and end of year testing is mandated, but more testing is conducted at a teacher's discretion);
- test results are returned promptly in a comprehensible and flexible format that matches teachers' planning needs;
- teachers receive training and support in how to use the system;

- the system receives strong and consistent support from school and district administrators.

The system described above provides one of many models that may be used to promote school improvement through the use of evaluation. Other models might permit greater variability and autonomy at local school sites, feature greater attention to the school and classroom processes which help to explain school effectiveness, assess outcomes beyond basic skills objectives, and consider a data set beyond student achievement data. Baker (1983) has described a more comprehensive, "top down, bottom up" school level system which includes demographic information, affective measures, indicators of school and classroom process, and district wide and local option measures of achievement -- a system that acknowledges the special and unique characteristics of each local school (including school problems, goals, processes and context). Sirotnik (1984) has described a similar model based on critical theory, one which emphasizes teachers' roles in the evaluation and inquiry process.

Multipurpose evaluation systems that are useful for local schools and that are used by them are possible. With the advent of micro-computers, schools can collect and easily make sense of a comprehensive data base in order to improve their decision-making and instructional programs. It is likewise feasible to aggregate and transmit relevant portions of that data base for school district decision-making and to disaggregate that information for teachers' classroom uses.

Whatever the model, it is important that the evaluation data get targeted at the point of where the information can get used to improve the instructional process. Targetting the information at this point, the ultimate point of change, means that we must design systems that teachers

actually will use, ones which will be a benefit rather than a burden. If we base these systems on what we know about local utilization (see, for example, Berman & McLaughlin, 1977), we will use a process which builds teacher ownership, provides flexibility and adaptability to teachers' needs, and insures strong leadership support at both the district and school levels. Teachers, in short, will need a strong role in the design, implementation, and control of the system, and in order to play this role, they may need staff development in evaluation skills and problem-solving orientations. This does not mean making teachers super-technicians or having them master the nuances of experimental design, instrumentation, t-tests, and the like. Rather it means providing them with conceptual tools for analyzing their school and classroom environments.

If evaluation is to meet its promise for school improvement, we need to stimulate schools and the staff within them to pursue a self-inquiry mode:

What are local goals and problems?

What are potential solutions, causes, strategies?

What data (including formal and informal varieties) can be collected to help understand and solve the problem/reach the goal?

How can the data be best analyzed and synthesized (what user friendly local computer options are available)?

What are appropriate next steps?

The model remains simple. Its implementation needs to acknowledge the uniqueness of the local school context, the importance of curriculum match, and the complexity of promoting change.

## REFERENCES

- Airasian, P.W. The effects of standardized testing and tests on teachers' perceptions and practices. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, California, 1979.
- Baker, E.L. Evaluating educational quality: A rational design. Invited paper at Center for Educational Policy and Management. University of Oregon, 1983.
- Bank, A., & Williams, R. Evaluation design project: School district organization study. Report to the National Institute of Education. Los Angeles, California: Center for the Study of Evaluation, 1981.
- Bank, A., & Williams, R. School district uses of testing and evaluation for instructional decision-making. CSE Report No. 204. Los Angeles, California: Center for the Study of Evaluation, 1983.
- Berman, P., & McLaughlin, M. Federal programs supporting educational change: Factors affecting implementation and continuation. Report R-1589/7-HEW. Santa Monica, California: Rand Corporation, 1978.
- Center for the Study of Evaluation. School test evaluations. Los Angeles, California, 1974, 1976.
- Center for the Study of Evaluation. Criterion-referenced test handbook. Los Angeles, California, 1979.
- Cohen, D.K., & Garet, M. Reforming educational policy with applied social research. Harvard Educational Review, 1975, 45(1), 17-43.
- Dorr-Bremme, D. Assessing students: Teachers' routine practices and reasoning. Evaluation Comment. Volume 6, No. 1. Los Angeles, California: Center for the Study of Evaluation, 1983.
- Dorr-Bremme, D., Burry, J., Catterall, J., Cabello, B., & Daniels, L. The costs of testing in American schools. CSE Report No. 198. Los Angeles, California: Center for the Study of Evaluation, 1983.
- Dorr-Bremme, D., Herman, J., & Doherty, W. Achievement testing in American public schools: A national perspective. In Herman, J. and Baker, E., project directors. Testing in the nation's schools: Collected papers. Report to the National Institute of Education. Los Angeles, California: Center for the Study of Evaluation, 1983.
- Gorth, W.P., & Perkins, M.R. A study of minimum competency testing programs: A final report. Washington, D.C.: Office of Testing, Assessment and Evaluation. National Institute of Education, 1979.
- Goslin, D.A. The use of standardized ability tests in American secondary schools and their impact on students, teachers, and administrators. New York: Russell Sage Foundation, 1965.

- Goslin, D.A., Epstein, R., & Hallock, B.A. The use of standardized tests in elementary schools. Second Technical Report. New York: Russell Sage Foundation, 1965.
- Hall, M. Statewide assessment of student performance: A comparative survey. Paper presented at the annual meeting of the American Educational Research Association, 1975.
- Herman, J. Serving the mutual needs of research and practice. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, 1984.
- Herman, J., & Cabello, B. Testing, evaluation and instruction: Problems in making them work together. Invited presentation to the Management Information Network. Los Angeles, California, 1984.
- Herman, J., & Dorr-Bremme, D. Uses of testing in the schools: A national profile. In W.E. Hathaway, Ed. Testing in the Schools. New directions for testing and measurement. No. 19. San Francisco: Jossey Bass, 1983.
- Huron Institute. Summary of the spring conference of the National Consortium on Testing. Cambridge, Mass.: Huron Institute, 1978.
- King, J., & Pechman, E.M. The process of evaluation in local school settings. New Orleans: 1982.
- Rudman, H.C., Kelly, J.L., Wanous, D.S., Mehrens, W.A., Clark, C.M., & Porter, A.C. Integrating assessment with instruction: A review 1922-1980. East Lansing Michigan: Institute for Research on Teaching, 1980.
- Salmon-Cox, L. Teachers and tests: What's really happening. Paper presented at the annual meeting of the American Educational Research Association. Boston, 1980. (Also Phi Delta Kappan, 1981, 631-634.)
- Sirotnik, K. Principles and practice of contextual appraisal for schools. Occasional Paper No. 5. Los Angeles, California: Laboratory in School and Community Education, 1984.
- Weiss, C.H. Utilization of evaluation: Toward comparative study. In C.H. Weiss (ed.), Evaluating action programs: Readings in social action and education. Boston: Allyn and Bacon, 1972.
- Worthen, B.R., & Sanders, J.R. Educational evaluation: Theory and practice. Worthington, Ohio: Charles A. Jones, 1973.
- Yeh, J. Test use in schools. Report to the National Institute of Education. Los Angeles, California: Center for the Study of Evaluation, 1978.

7