DOCUMENT RESUME

ED 267 501 EA 018 298

AUTHOR Fielding, Glen; And Others
TITLE Effects on Teacher Practice of a Staff Development Program for Integrating Teaching and Testing in High School Courses. Final Report.
INSTITUTION Oregon Univ., Eugene. Center for Educational Policy and Management.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Mar 86
NOTE 73p.
AVAILABLE FROM Publication Sales, Center for Educational Policy and Management, University of Oregon, Eugene, OR 97403 ($3.00 prepaid; quantity discounts; postage will be added to billed orders).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Faculty Development; High Schools; Inservice Teacher Education; *Instructional Improvement; Secondary School Teachers; Teacher Effectiveness; Teacher Made Tests; Teaching Methods; Test Construction; *Testing; *Test Use

ABSTRACT

An attempt to improve high school teachers' integration of testing and instructional practice through inservice training failed to produce substantial results. In each of five districts participating, one high school used the intervention and one served as a control. Three science teachers were trained in each school, one as a lead teacher. Data were gathered both before and after the training period in three forms: classroom observations of teachers' practices and teacher and student questionnaires. The data generated failed to support the hypotheses that trained teachers would prove more effective than untrained teachers in (1) communicating learning goals; (2) communicating what, how, and when learning is to be assessed; (3) using tests reflecting learning goals; (4) using a variety of test formats; (5) establishing criteria for evaluating responses; or (6) providing specific feedback to students. The seventh hypothesis held that frequent use of information from unit tests would serve as a guide for planning corrective or enriching instruction. A few teachers increased their use of test results in planning correctional instruction, thus supporting the seventh hypothesis. This report describes the sample, the research methodology, and the findings, including teacher assessment of the training program's effectiveness. Directions for future efforts are suggested. (PGD)

Effects on Teacher Practice of a Staff Development Program for Integrating Teaching and Testing in High School Courses

by
Glen Fielding, Joan Shaughnessy,
and Kenneth Duckworth

March 1986

Center for Educational Policy and Management
College of Education
University of Oregon
Eugene, Oregon 97403

The preparation of this final report was made possible through an Institutional Grant awarded by the National Institute of Education to the Center for Educational Policy and Management. The opinions expressed in this report do not necessarily reflect the positions or policies of the NIE or the U.S. Department of Education.

ACKNOWLEDGMENTS

We would like to thank our colleagues Del Schalock, Mark Gall, and Sandy Charters for their help in developing the research methodology used in the study. We also extend our thanks to Berk Moss and Lou Mueller, from the Beaverton School District, for assistance in planning the staff development program. And we are grateful to Ron Renchler for generous editorial assistance.

Perhaps most of all, we are indebted to the teachers and administrators who graciously participated in the study. We hope our findings will be of value to them and the schools in which they work.

CONTENTS

Page

CONTENTS, continued

APPENDICES

Appendices referred to in this report can be obtained for $3.50 from the Center for Educational Policy and Management, College of Education, University of Oregon, Eugene, Oregon 97403.

CHAPTER ONE

BACKGROUND AND PURPOSE

Introduction

This report deals with the implementation and impact of a staff development program for high school teachers on integrating teaching and testing. The program aimed to increase teachers' skills in designing and using tests to facilitate instruction and aid student learning.

This study is part of a larger study on high school teachers' design and use of tests. The project is sponsored by the Center for Educational Policy and Management at the University of Oregon, through funds provided by the NIE. In addition to the intervention described in this report, the larger project includes a study of the relationship between test-related practices and students' academic effort and feelings of academic efficacy. Results from this companion study will be presented in a forthcoming report.

Why a Training Program on Integrating Teaching and Testing?

The rationale for focusing on the linkage between instruction and assessment in a staff development program for high school teachers was derived from four sources. The first was the growing body of literature on mastery learning (Bloom, 1976; Guskey, 1985; Ryan and Schmidt, 1979). Mastery learning is an instructional model that calls for clarity about the learning outcomes expected from instruction, the use of "formative" tests to provide information for both students and teachers on students' progress toward outcome attainment, and the use of "corrective" instruction for students whose progress is unsatisfactory and "enriching" instruction for those who master material quickly. Reviews of research studies and evaluations of mastery learning programs offer firm evidence that mastery learning can assist teachers in improving student learning at a variety of educational levels and in a variety of subject areas (Block and Burns, 1976; Fitzpatrick, 1985; Guskey and Gates 1985; Ryan and Schmidt, 1979).

A related line of inquiry that suggested the importance of integrating teaching and testing was the research on the characteristics of effective schools (Austin, 1979; Cohen, 1982; Rutter et al., 1979). As Porter (1983) has noted, nearly all reviews of the effective schools literature indicate that high-performing schools have in place systems for assessing pupil performance in reference to established learning goals and for managing and using assessment information. In essence, the effective schools literature suggests an image of schooling that "bears a marked resemblance to a mastery learning model, but at the school level" (Porter, 1983, p. 26). Although the staff development program dealt with in this report focused on instructional units within individual courses rather than on the school as a whole, a focus on units and courses seems to be an appropriate starting point for efforts to develop and implement schoolwide systems for assessing students' learning and using assessment information.

Another source of support for the inservice program was research on the effect of evaluation on students' attitudes toward schoolwork and toward themselves as learners. There is some evidence that high school students respond to academic work in line with expectancy theory, in which effort is a function of (1) the valuation of rewards attendant on successful performance and (2) the perception that effort will lead to successful performance (Lawler, 1976). Natriello and Scott (1981) found that substantial numbers of high school students perceive that tests are not good indicators of learning and that hard work does not always lead to satisfactory test performance. It appears that, in high schools, increased emphasis needs to be placed on helping students to see connections between the focus and form of coursework and assignments on the one hand and tests and performance standards on the other.

Integrating teaching and testing also seemed to be a worthy target for a staff development program in view of the mounting evidence that teachers (1) have particular difficulty in evaluating student learning and using evaluation results and (2) receive little training in this area. Barnes (1985), for example, examined the attitudes, conceptions, and practices of cooperating teachers and student teachers at both the elementary and secondary levels with respect to the evaluation of students' learning progress. Both cooperating and student teachers indicated that evaluation and grading were the most difficult aspects of teaching. Neither group articulated clear criteria for assessing student learning or for translating assessment results into grades. Both groups seemed to rely on informal observations of and oral interactions with students to evaluate what and how well they were learning. Teachers were apt to assign grades to students on the basis of intuitive judgments about "where students were" and how hard they had worked. With the notable exception of one student teacher who was working in a school that had adopted mastery learning on a schoolwide basis, none of the teachers in the study appeared to have developed a systematic approach to assessment and the use of assessment information. Although Barnes did not focus on differences between elementary and secondary school teachers in her sample, recent studies of instruction in high school suggest that teachers at grades 9-12 are even less likely to use tests to guide and improve instruction than are elementary school teachers (Farrar, Neufield, and Miles, 1984).

Finally, research on the preparation and continuing education of teachers indicates that teachers typically receive only a small amount of training in evaluation and the use of evaluation findings, and that high school teachers receive even less training than elementary school teachers (Stiggins and Bridgeford, 1982). To the extent that staff development programs in this area are available, they generally focus on administering and interpreting tests mandated and managed by either the school district or the state (Burry et al., 1982). Without specific training, teachers are likely to lack not only a conception of evaluation as a systematic process, but technical skills in assessment as well, such as the ability to determine the degree to which a given set of test items corresponds to a given learning goal (Carter, 1984). The rationale for conducting an intervention in secondary schools around the theme of integrating teaching and testing thus seemed clear and compelling.

Development of the Inservice Program

The program implemented in this project built upon materials prepared through a related contract with the NIE (No. 400-82-0013). The purpose of this earlier project was to develop (1) a handbook for high school teachers on integrating teaching and testing and (2) guidelines for organizing and carrying out an inservice program based on material included in the handbook. Detailed information about the work accomplished through the development project and the characteristics of the products produced is contained in a technical report (Fielding and Schalock, 1985a). The products themselves currently are available through the Teaching Research Division of the Oregon State System of Higher Education.

The current study extended the original project in two ways. First, whereas the earlier project was intended to develop inservice training materials, this project involved research on the effects of training. More specifically, the current study investigated the effects of the "basic" training program that had been developed in draft form under the previous contract. (The basic program focuses on linking instruction and assessment in individual instructional units. The "advanced" program deals with teaching and testing in the context of a course as a whole).

A second major difference between the projects was that the earlier one focused on all subject areas, whereas the present study focused on science alone. Science was selected as the focal content area because of widespread interest at both national and regional levels in instructional improvement in this area. Science also was singled out because of the wide range of learning outcomes associated with scientific disciplines. Such outcomes include mechanical skills (e.g., using a microscope or Bunsen burner properly), inquiry skills (e.g., designing meaningful experiments), and problem-solving skills (e.g., determining how best to eliminate pollution from a local stream). Learning outcomes in science lend themselves to a wide variety of testing formats and procedures.

This decision to focus on science carried with it, the researchers felt, an obligation to develop specific materials in science over and above the general training materials prepared through the original development project. Accordingly, a professor of science education with an extensive background in the area of assessment was asked to serve as a consultant to the project and to develop test items and other assessment procedures that illustrate how different types and levels of learning outcomes in science could be measured. This collection of illustrative assessment tools is contained in Appendix A.

Hypotheses

It was hypothesized that, compared to nonparticipating teachers, teachers participating in the staff development program would:

1. Communicate more clearly to students the learning goals they are expected to achieve in particular instructional units;
2. Communicate more clearly to students what, how, and when learning is to be assessed;
3. Use tests that reflect more closely the learning goals they have established;
4. Use a greater variety of test formats, e.g., essay questions and open-ended, problem-solving items in addition to multiple choice, true-false, and matching items;
5. Establish clearer criteria for evaluating student responses to essay questions and other items that require judgment to score;
6. Provide more specific feedback to students on what they have and have not learned; and
7. Make more frequent use of information from unit tests as a guide to planning corrective or enriching instruction, either for the class as a whole or for individuals or groups within the class.

CHAPTER TWO

RESEARCH METHODOLOGY

Research Design

The effects of the intervention were investigated using a pre-post experimental-control group design. Two high schools from each of five districts were nonrandomly assigned to either a staff development or control condition. The research design can be shown as follows:

O_1, O_2, O_3, X_1, O_1, O_2, O_3

O_1, O_2, O_3, X_2, O_1, O_2, O_3

where:

O_1 = Observational measure of teachers' test-related practices

O_2 = Teacher questionnaire on test-related practices

O_3 = Student questionnaire on test-related practices

X_1 = Inservice training program

X_2 = Control condition; no training provided

Three science teachers from one high school in each participating district were assigned to the staff development condition; three science teachers from another high school in the district were assigned to the control condition. Two of the teachers from each school taught biology; one taught either physical science or chemistry. Each teacher was asked to select one particular course and class period to serve as a context for data collection during the study.

The sampling plan for the study is shown in Table 2-1.

Recruitment of Sample

The recruitment process was carried out in the spring and summer of 1984. Project staff first identified nine school districts in western Oregon and southwestern Washington that appeared to have some interest in strengthening the integration of

Table 2-1

The Sampling Plan for Research on the Staff Development Program

	Biology Classes	Chemistry or Physical Science Classes
Control Condition		
School 1	N = 2	N = 1
School 2	N = 2	N = 1
School 3	N = 2	N = 1
School 4	N = 2	N = 1
School 5	N = 2	N = 1
	10 classes	5 classes
Staff Development Condition		
School 6	N = 2	N = 1
School 7	N = 2	N = 1
School 8	N = 2	N = 1
School 9	N = 2	N = 1
School 10	N = 2	N = 1
	10 classes	5 classes
TOTAL	20 classes	10 classes

teaching and testing. The project staff had prior experience in working with these districts and had reason to believe that they would be receptive to the proposed inservice program. The research design required a sample of only five districts, but nine were contacted under the assumption that not all districts would agree to participate. Following presentations to district and school staffs, two high schools in each of five districts chose to participate.

Although central office staff often played a key role in deciding whether a school would participate in a project, it was high school principals who decided whether the school would serve as either a treatment or control site. Some principals clearly made their decision on the basis of input from science teachers in the school. Others appeared to have made their decision before consulting with staff. Consequently, in some schools, the teaching staff supported the project from the start. In a few schools, however, the staff initially was somewhat resistant to an instructional improvement project not of its own choosing.

Once a school made a decision to participate, a "lead" teacher and two "regular" teachers had to be selected. In all schools, principals permitted the science department staff to develop its own procedures for choosing a lead and regular teachers. Informal reports suggest that in at least two departments a bit of good-natured arm-twisting was involved in finding a volunteer for the position of lead teacher. Volunteers for the roles of regular teachers generally were easier to find.

Description of Sample

Characteristics of the districts, schools, and teachers participating in the study are summarized in the pages that follow.

District and School Characteristics

Information on each school in the sample was collected in the spring and fall of 1984 through interviews with building administrators and science department chairs.

A summary of this information is presented below. It focuses on the characteristics of the community and students served by each school, administrative policies pertaining to student assessment and grading, and recent staff development efforts in which the science department was involved.

Before focusing on individual districts and schools, it should be noted that the sites had much in common. All ten schools were located in the Willamette Valley in western Oregon. All served primarily a white, middle-class population. All of the schools offered a traditional science curriculum, emphasizing college preparatory coursework in biology, chemistry, and physics. Although instructional improvement projects were underway or were being planned in several districts, administrators generally regarded the quality of instruction in their buildings to be high. None of the schools was "in trouble"; all seemed to be reasonably well-managed institutions.

District A. District A was in a medium-sized city. The city's population included working-class families and middle-class professionals. The enrollment in the two schools in the study ranged between 1,000 and 1,300 students. Both schools served grades 9-12, although 1984-85 was the first year in which the treatment school included a ninth grade.

With respect to testing and grading policies, the district was on a trimester system under which teachers were required to provide six grade reports during the year, in contrast to the four reports required under the more conventional semester system. The district required teachers to prepare written statements of their grading procedures for students. However, no explicit guidelines for doing this had been established.

The district was entertaining the idea of adopting a long-term staff development program to foster the improvement of instructional skills, but the program had not yet been implemented.

District B. This district was in a suburb of a large city. Its population consisted mainly of middle-class to upper middle-class families. The district had a statewide reputation for academic excellence. The two high schools participating in the study both had enrollments of over 1,500 students in grades 10 to 12.

The district had been involved in an extensive staff development effort designed to promote teachers' use of a model of instruction called "Elements of Effective Instruction," commonly referred to as "Instructional Theory into Practice," or ITIP (Hunter, 1976). This model is complementary to the inservice program that was implemented in the present study in that ITIP emphasizes the importance of being clear about learning objectives, communicating objectives to students, monitoring students' learning progress, and adjusting instruction in view of how well students are learning. ITIP, however, focuses primarily on instruction and informal assessment carried out in the context of individual lessons, whereas the inservice program implemented in this study focused on formal assessment of the learning accomplished over broader segments of instruction and on the uses to be made of results from these assessments. The building administrators in both the treatment and control schools in this district were optimistic that the inservice program on integrating teaching and testing would complement the ITIP program already in place. In fact, in the treatment school, the principal used district funds to involve more science teachers in the program than the budget for the study could support.

With respect to testing and grading policies, both schools required that final examinations be given in each course. Each also operated under a school policy which specified what percent of a student's grade could be based on scores from the final.

District C. This district served a heterogeneous population in a medium-sized city. About 1,500 students were enrolled in each of the two schools participating in the study.

The district was making an effort to standardize the course offerings in its high schools. Course offerings for all schools were listed in one catalogue.

In the other districts in the sample, each of the high schools published its own catalogue of courses.

The district also had developed specific policies on grading, including a directive that assessment must be based upon the "goals, objectives, and scope and sequence described in the Planned Course Statement." Teacher expectations for grades were to be consistent across multiple sections of the same course, even if several teachers taught that course.

Although many teachers in the district had participated in an ITIP inservice program, the science departments of the schools participating in this study had not been systematically involved in this program. The district coordinator of science education indicated that more and more teachers would be participating in ITIP. Like the building administrators in District B, he thought that the program on integrating teaching and testing would complement the ITIP program.

District D. This district was also located in a city of moderate size, the residents of which were predominantly professional in background. The students in the district had a reputation for being high-achievers. Each of the high schools enrolled about 1,000 students.

There were no district or school policies that related specifically to testing or grading, nor had any staff development programs been carried out recently with the science department staff. In the treatment school, however, the principal, who had been in his post for only one year when the study began, was attempting to initiate instructional improvement efforts. He viewed the inservice program associated with this study as supportive of the broader school improvement plans he was developing. For example, he hoped that the school-based workshops would result in greater collaboration within the science department.

District E. Like District B, this district was located outside of a large city. Although the majority of students were from suburban, middle-class areas, a sizeable minority came from rural sections of the community. Each high school served about 1,500 pupils.

The two high schools in the district appeared to operate with a high degree of independence from each other and from the district office. In the treatment school, administrators indicated that particular emphasis was placed on teachers' development of Planned Course Statements, which were expected to guide instruction and assessment in each class. Teachers were required to update these statements for all their courses each year. They also were expected to state their grading practices in writing and to submit them to their division leader, a position analogous to a department head. Absences and latenesses were not to figure into the assignment of course grades.

The treatment school recently had been involved in a staff development program dealing with student learning styles, but there seemed to be widespread teacher dissatisfaction with that project. Administrators viewed the inservice program to be carried out in this study as holding particular promise. As in District B, the administration decided to include, at district expense, two more teachers in the program than project staff had budgeted for.

Administrators in the control school in District E did not appear to place as much emphasis on planned course statements as did administrators in the treatment school. However, the school did publish a teacher's handbook that provided suggested guidelines for grading on a curve and that outlined a mandatory set of procedures for maintaining a gradebook. Administrators in this school were interested in test improvement and had considered the development of a criterion-referenced assessment program to evaluate student learning in various curriculum areas, including science. Administrators also had encouraged teachers to place greater attention on teaching and assessing higher-order thinking. In connection with this thrust, the school planned to sponsor an inservice workshop on test improvement in January, 1985, but this never materialized.

Teacher Characteristics

Information on the background and working conditions of teachers in the sample was obtained from a teacher questionnaire administered in November 1984. The sample consisted of 22 males and 8 females. (Females were equally divided between the two experimental conditions.) The sample was racially homogeneous; only one teacher in the entire sample was nonwhite.

The variables related to working conditions that were of primary interest were the number of different courses a teacher taught and had to prepare for each day, the total number of students in a teacher's classes, and the time allocated to a teacher each day for instructional planning. Information related to these variables is shown in Table 2-2. (The information was obtained from the teacher questionnaire, which is discussed later in this chapter.) Also included in the table is information on the number of years teachers in the sample had been employed in the school in which they currently worked.

It is noteworthy that in four of the five participating districts, teachers in the treatment school had between five and twenty minutes less planning time each day than teachers in the control condition. It is unclear why this is so; it may well be a mere coincidence.

Training and Support of Lead Teachers

A noteworthy characteristic of the inservice program implemented in this study is that it involved two levels of training, one for "lead" teachers and the other for "regular" teachers. The role of lead teacher was intended to be filled by teachers who had special interest in classroom assessment and the use of assessment information and who had a commitment to increasing their skills in this area. Lead teachers, furthermore, needed to be willing to work collaboratively with colleagues and administrators on instructional improvement matters.

The idea of establishing and training lead teachers, who would in turn train and support their colleagues, appeared to reflect a growing and seemingly productive trend

Table 2-2

Descriptive Information about Treatment and Control Groups of Teachers

District	Treatment or Control	Mean No. of Years in This School	Mean No. of Separate Preps.	Mean No. of Students Taught Each Day	Mean No. of Minutes of Planning Time Per Day
A	T	1.3	1.7	118	50
	C	7.3	1.7	118	70
B	T	11.0	2.0	142	50
	C	4.3	2.0	126	50
C	T	5.3	2.7	134	45
	C	13.7	2.3	133	55
D	T	7.0	2.3	146	50
	C	6.7	2.0	133	65
E	T	9.0	2.3	117	45
	C	10.7	2.3	131	50
All Treatment		6.5	2.2	131	48
All Control		8.5	2.1	128	58

in inservice education. Recent research suggested that lead teachers can play key roles in school improvement projects (Hord et al., 1984). The past experience of one of the researchers in working with lead teachers at the elementary school level (Fielding and Schalock, 1985b) also suggested that lead teachers could be a valuable resource in staff development efforts. Lead teachers can be highly effective in translating the often abstract language and ideas of researchers and external change agents into terms that are meaningful for particular groups of teachers in particular school settings (Beaton, 1985).

As indicated earlier, one teacher from each school in the treatment group served in the role of lead teacher. A two-day training program for lead teachers was held in late November 1984. The training was guided by a goal-based approach to integrating teaching and testing. Teachers received assistance in clarifying and upgrading the learning goals in science they expected students to achieve; developing tests that corresponded tightly to established goals; providing feedback to students on their progress toward goal attainment; and deriving instructional implications from test results. In addition to having an opportunity to discuss these practices and the rationale underlying them, teachers were given the chance to apply what they were learning to a specific instructional unit of their own choosing. It was intended that by the end of the two-day program participants would be able to produce in draft form: (1) learning goals for a unit they planned on teaching later in the year, (2) a test that could be used to assess goal attainment, (3) appropriately demanding performance standards for the test, and (4) a plan for scoring the test and for analyzing, reporting, and acting upon test results.

Lead teachers also received guidelines and materials for carrying out a similar training program with colleagues in their home districts. Recommended agendas for the school-based training sessions were distributed and discussed. These are presented in Appendix B.

The principal or assistant principal of each of the lead teachers' schools attended the second day of the training program. This was done so that he or she could (1) gain an understanding of the program and the approach to integrating teaching and testing with which it dealt and (2) explore ways of helping lead teachers implement the program in their departments.

Following the training session, project staff met individually with each lead teacher to provide further preparation for the school-based training that the lead teacher would be facilitating later in the school year. The school-based training was to parallel in focus and format the training that the lead teachers received. The role that project staff played in the workshops for lead teachers represented a model of the role that lead teachers were to play in the school-based workshops. However, given the small amount of training that the lead teachers received, it was considered unreasonable to expect them to conduct inservice sessions with their colleagues without any external assistance. Consequently, a project staff member was to attend each workshop and serve as a resource for the lead teacher.

One of the main tasks accomplished during the planning meetings between lead teachers and project staff was to gain increased clarity about the kind of support to be offered by the staff. Lead teachers were well aware that they would be primarily responsible for organizing and conducting the training, but it also was important to establish that project staff would lend a hand if and when it was needed or requested.

Training and Support of Regular Teachers

Training for regular teachers followed the same overall agenda and focused on the same objectives as training for lead teachers. The workshops took place between December 1984 and February 1985. In three of the five schools in the treatment group, the workshops were held in the school building. Teachers in the remaining two treatment schools chose to meet away from the school site for at least a portion of the inservice training.

A project staff member attended each of the training sessions to provide support to lead teachers. The staff member contributed some ideas and suggestions at each workshop, but lead teachers carried the main burden for facilitating the discussions and activities.

Following the training sessions, several support activities were to take place. By mid-March, an informal conference was to be held between the lead teacher and regular teachers in each participating school to review plans for implementing new or refined practices. By the end of May, meetings were to be conducted between lead teachers and building administrators, and another meeting was to be held between lead teachers and regular teachers. A final "sharing" of products developed and lessons learned through the project was to occur by the close of school in June. The complete schedule for the staff development program is presented in Table 2-3.

Role of Building Administrators

Building administrators were asked to attend a portion of the inservice workshops to convey to participating teachers that the school stood behind the inservice project and viewed its objectives as important. Administrators also were expected to indicate how the project related to the school's current or recent instructional improvement efforts. In addition, it was anticipated that building administrators would provide encouragement and furnish lead teachers with informal support to help them carry out their roles in the project. Finally, building administrators were asked to discuss with participating teachers substantive issues that might arise during the course of the project, for example, how a mastery-learning approach could be implemented effectively in view of the large scope of content that most high school science teachers are expected to cover in all classes.

Specific tasks to be carried out by building administrators are listed in Table 2-4. Whether a principal decided to carry out these tasks personally or to delegate them

Table 2-3

A Schedule for the Staff Development Program

November 28-29, 1984	. Leadership Conference on Integrating Teaching and Testing November 28: Lead teachers only November 29: Lead teachers and building administrators
by January 10, 1985	. Individual conferences between lead teachers and a project staff member to prepare for the school-based work sessions. . Individual conferences between lead teachers and building administrators to finalize plans for the work sessions and follow-up activities.
by February 28, 1985	. Two full-day teacher work sessions (the second to take place between 3 and 7 days after the first session), facilitated by lead teachers. These are to take place in each participating school. A project staff member will attend both sessions.
by March 15, 1985	. A conference between lead teachers and their department colleagues who are participating in the project to review plans for implementing new or refined practices.
between March 18 and May 24	. A meeting between lead teachers and building administrators to discuss progress that has been made on the project and issues that have arisen. . A conference between lead teachers and department colleagues to discuss the implementation and perceived impact of designated practices.
by the end of the school year	. Lead teachers and department colleagues share products developed and lessons learned through the project with others in the department, school, or larger professional community.

Table 2-4

Tasks for Building Administrators

Date	Task	Time Requirements (Estimates)
November 29, 1984	Attend the second day of a two-day inservice session on classroom testing and on strategies building administrators can use to foster effective student evaluation practices.	5 hours (in addition to driving time)
by January 10, 1986	Meet with your building's lead teacher who will be facilitating the school-based portion of the inservice program to finalize plans for program implementation.	20 minutes
by February 28, 1985	Attend at least 20 minutes of each of the two school-based inservice sessions.	40 minutes
between April 1 and May 3, 1985	meet at least once with the teachers participating in the project to discuss progress that is being made in implementing new or refined practices and to resolve issues that may have arisen.	1 hour
	Share your perception of the project and your role in it with a member of the project staff.	30 minutes
	Total:	7 ½ hours

to an assistant principal was left totally to the principal's discretion. In three of the schools, an assistant principal was assigned project-related responsibilities. In one school, the principal shared responsibilities with an assistant. In another school, the principal carried the load himself.

Observational Measure of Test-Rela.ed Practices

Project staff developed procedure for observing classrooms when teachers were reporting results from unit tests to students and responding to these results in class. This observational measure focused primarily on the nature of the feedback a teacher gave a class on its test performance and the instructional response a teacher made to test information. Supplementing the observation were guidelines for conducting brief interviews with teachers. These interviews were intended to provide information about the context in which an observation took place - for example, the nature of the content a particular unit test covered, the procedures a teacher used to design or select the test, and whether a teacher made written comments on test papers to supplement scores or grades. Also, during the brief interview, observers asked to see a copy of the teacher's test. This helped the observers to make sense of classroom discussion of test results and obtain accurate information on the characteristics of the test. For the sake of simplicity, the term "observational measure" is used to refer both to the observation form and the related interview procedure. A copy of the observational measure is presented in Appendix C.

Development, Pilot Testing, and Observer Training

The project staff prepared the initial draft of the observational measure in March 1984. It was refined in April and then refined further in May and early June during training sessions with the four individuals hired to assist project staff in carrying out the observations. After each of the five observer training sessions, revisions were made in the instrument. These revisions provided clearer and more specific descriptions of the practices to be observed; widened the range of practices to be observed, particularly with respect to the category, "Instructional response to test results"; and broke the observation into two phases: a) maintenance of a running record of the "flow" of events and statements and b) translation of the running record into a coding system.

During the course of training, the instrument was used as a guide for observing three videotaped lessons, one audiotaped lesson, and three lessons as they were being carried out in the classroom. The three classroom observations were conducted in biology and chemistry classes. The taped lessons were taught by two science teachers, a foreign language teacher, and a social studies teacher.

Interobserver reliability in using the instrument was high during the last two training sessions; discrepancies were found only with respect to two observational categories. To reduce these discrepancies, the measure was revised during the summer.

In September 1984, the instrument was submitted for review to a technical panel that the project directors had established to provide advice on issues of measurement and data analysis. The panel consisted of Drs. Sandy Charters and Mark Gall, from the University of Oregon, and Dr. Del Schalock, from the Teaching Research Division, Oregon State System of Higher Education. The panel suggested several additions to the section on scoring tests and to the postobservation interview. These additions were made.

The final version of the instrument was field-tested in late September. The four members of the observation team and two project staff members used the measure when observing a videotape of a class discussion of results from a unit test. A high level of interrater agreement was achieved. Nonetheless, it was decided that a project staff member would accompany observers on their first classroom visit to ensure that the instrument was sensitive to classroom variability and practical to use under actual field conditions.

Use of the Measure

Of the 30 teachers participating in the study, 29 were observed twice before training, in October and November, and twice after training, between March and May. One teacher was observed three times rather than four because he did not make clear to the observers his plans for giving tests and passing back test papers to the class. When the observers attempted to conduct the fourth and final observation, they discovered that he already had handed back to the class the last unit test of the year. Observers were not informed which school had been assigned to the staff development condition and which to the control condition.

Reliability

To determine interrater reliability, a member of the project staff accompanied observers on 18 (15%) of the 119 classroom visits. Five of these reliability checks were conducted during the observer's first classroom visit. The staff member and the observer used separate observation forms to record data from the visit. After completing the observation, the two forms were compared. Information on the degree of consistency among raters is summarized in Table 2-5.

Overall, interrater agreement was high. There was 90% agreement or higher on 19 of the 27 variables for which reliability statistics were calculated. On 14 of these 19 variables, agreement was 100%.

Lower levels of interrater consistency occurred on items that called for more subjective judgment on the observer's part. For example, observers and staff disagreed 4 times out of 18 in rating the level of student involvement in the lesson observed, although only once did ratings differ by more than 1 point on a 4-point scale.

Teacher Questionnaire on Test-Related Practices

A questionnaire was developed to obtain teacher-reported data on the type and frequency of test-related practices used in the classrooms in the sample. The

Table 2-5

Interrater Agreement in Using the Observational Measure

Focus of Observation and Supporting Interview[a]	Percent of Interrater Agreement
Were test items based on written goals?	83
Was test self-made, included in curriculum materials, other?	100
What types of items were used?	
. Multiple-choice	94
. True/false	100
. Matching	100
. Fill-in-the-blank	100
. Short answers	94
. Open-ended problem solving or essays	[b]
Did teacher calculate a total test score? subscores? both?	94
Did teacher make written comments on test papers?	100
. Percentage of papers containing comments	100
. Type of comments	83
Type and specificity of teacher report to class on its test performance	100
If results reported for some items and not others, did teacher explain why?	100
Extent to which teacher verbally stated (or had students state) correct answers for:	
. Response-selection items (e.g., multiple-choice, true/false)	89
. Response-completion items (e.g., fill-in-the-blank, short answer)	83
Extent to which teacher explained why particular responses were correct/incorrect:	
. Response-selection items	78
. Response-completion items	89
If partial credit given, teacher explained basis for awarding credit	95
Extent to which students raised questions as to clarity or fairness of items	95

Table 2-5, continued

Focus of Observation and Supporting Interview	Percent of Interrater Agreement
Number of items questioned by students that teacher decided not to count	100
Teacher's approach to scoring and reviewing results from essay questions or open-ended problem-solving items	b
Extent to which teacher focused on students' underlying misconceptions	89
Teacher conducted review of test results with whole class? small groups? combination?	100
Special work given to class as a whole based on test results?	b
Special work given to individuals or groups based on test results?	b
General level of student involvement in the lesson observed	78
Teacher's perception of students' ability	100
Teacher's satisfaction with the lesson observed	100
Teacher's plans for next day's lesson	100
Teacher's plans for giving make-up tests for absent students	100

[a] For ease of reading, the variables listed in the table appear in a different order than they appear on the actual instrument (Appendix C).

[b] This practice was not observed in the class periods sampled for reliability testing.

questionnaire was administered in the fall, shortly before training, and in the late spring, after training. This measure dealt with issues related to:

(a) The communication of test expectations (e.g., frequency with which teachers inform students at the beginning of a unit what they will be expected to know on the unit test);

(b) The design of tests (e.g., frequency with which teachers use a written list of learning goals as a guide to developing or selecting test items);

(c) The scoring of tests and the reporting of test results (e.g., frequency with which teachers report information on student performance in reference to specific learning goals or areas); and

(d) The type of instructional responses that are made to test information (e.g., frequency with which teachers arrange peer tutoring for students who do poorly on tests).

The questionnaire also dealt with variables related to teachers' backgrounds, the conditions under which they worked, and the degree of collaboration that existed in their departments. A copy of the questionnaire is presented in Appendix D.

The questionnaire was pilot tested in the spring of 1984 among five science teachers. Items that teachers found unclear or difficult to respond to were revised. Also revised or eliminated were items that failed to discriminate among teachers. The questionnaire was then reviewed by the technical panel referred to earlier, and fine-tuned in light of the panel's critique.

The teacher questionnaire addressed many of the same variables as the observational measure of test-related practices discussed earlier. The observational measure, however, focused on a specific unit test, whereas the questionnaire focused on test-related practices teachers used throughout the year.

Student Questionnaire on Test-Related Practices

An additional questionnaire was developed to assess student perceptions of test-related practices, as well as their level of academic effort (how hard they worked in a course) and their feelings of academic efficacy (the extent to which they believed that hard work would lead to successful performance in a course). The questionnaire also included items on students' academic background and aspirations. For purposes of this study, the items of primary interest on the student questionnaire were those that elicited student reports of teachers' communication of test expectations, their procedures for giving feedback to students on their test performance, and their responsiveness to students' learning problems. The relationship between student responses to these items and their report of academic effort and efficacy will be examined in a forthcoming report.

The student questionnaire was pilot tested in the spring of 1984 with the students whose teachers participated in the pilot test of the teacher questionnaire. A total of 133 students completed the pilot test version. Several items failed to show satisfactory variation; these were revised. For example, the item "When I do well on tests in this class, it is mostly due to luck" produced little variation in student response. It was therefore decided to delete the word "mostly" on the final version of the instrument. Additional refinements were made on the basis of suggestions made by the technical panel.

The student questionnaire was administered a pre-post treatment basis, on the same dates as the teacher questionnaire. This measure is presented in Appendix E.

Development of Impact Scales

The observational measure and the two questionnaires discussed above assessed a wide range of variables, not all of which related directly to the teaching-testing practices dealt with in the staff development program. For example, the observational measure yielded information on how teachers dealt with students who were absent the day tests were administered. This subject was never discussed in any of the training sessions, and it clearly was outside the area of expected impact of the intervention. Information was collected on this subject as part of an effort to gain a comprehensive picture of test-related practices, not as a basis for testing the effects of the inservice activities. The information obtained about testing practices unrelated to the training will be analyzed and summarized in a forthcoming report.

In order to determine the impact of the program, those items on each of the three measures that pertained explicitly to the objectives of the program were identified. Patterns of response to the items identified were then examined to verify that the items were valid indicators of the practices in question. Several items from the measures were eliminated at this point because responses to them produced contradictory or highly inconsistent patterns. For example, one item on the observational measure and one item on the teacher questionnaire dealt with teachers' use of subscores on tests. However, a zero correlation was found between these two items on the pretreatment measures and a negative correlation was found on the posttreatment measures. Both items were designed to measure the same practice, but they elicited contradictory responses, and there was no way to determine which was the better measure. Both items therefore were eliminated.

Once the validity of the items was checked, "impact scales" were created from the observational measure, the teacher questionnaire, and the student questionnaire. The impact scales consisted of those items from each measure that pertained explicitly to the practices dealt with in the staff development program and met the assumptions about validity discussed above.

Table 2-6 lists the items from the observation instrument used to form an impact scale and the range of points that a teacher could receive on each. These values were derived by summing the scores that could be obtained from each of the two observations

Table 2-6

Items from the Observation Instrument Used in Constructing the Impact Scale

Item	Point Value	Variable Measured
3b	0-2	Unit learning goals displayed, read to, or shared with students through class handouts
3a	0-2	Test based upon written statements of learning goals
1f	0-2	Essay questions or open ended problems comprised 10% or more of total point value of unit test
II.1	0-2	Teacher scored essay questions guided by rating scales, explicit criteria, exemplary responses or other clear standards of quality
2c	0-2	Teacher provided written feedback directly on student test papers
I.1	0-2	When reporting results to class, teacher referred to separate sections of tests
III.2	0-2	Teacher focused reviews on student misconceptions
III.3	0-2	Special work given to class based upon students' test performance
III.4	0-2	Special work given to individuals or small groups based upon students' test performance
	—	
	0-18	(Range of points possible)

that were conducted before and after training. Put differently, the point values listed are double what could be earned through any one observation.

Table 2-7 lists the items from the teacher questionnaire used to form an impact scale. Items from the student questionnaire used to construct an impact scale are shown in Table 2-8.

Correlations among the impact scales both before and after training are presented in Table 2-9 and 2-10. The correlation between the observational measure and the teacher questionnaire was statistically significant before training, and in the treatment group, after training as well. Correlations between the teacher questionnaire and the student questionnaire and between the observational measure and the student questionnaire were consistently low. This suggests that the students tended to perceive classroom teaching and testing practices from a different perspective than the teachers and observers. The weak correlations might also reflect differences in focus and emphasis among the measures. For example, the observational measure focused on practices related to four specific unit tests, whereas the student questionnaire asked about a teacher's testing practices in general.

Correlations between scores from each observation conducted before and after training are shown in Table 2-11. With the exception of the correlations related to the item on teachers' use of written feedback, the correlations were moderate to low. This suggests that teachers' test-related practices may vary from one month to another. Whether this variation takes place to accommodate different types of subject matter or varying student needs, or whether the variation simply reflects inconsistency on the part of teachers is unclear.

Development of Measures for Each Hypothesis

The impact scales were intended to provide information on the level of classroom use of targeted teaching-testing practices as a whole. The scores did not indicate how well specific practices were being used.

Table 2-7

Items from the Teacher Questionnaire Used in Constructing the Impact Scale

Item	Point Value	Variable Measured
22c	1-3	Uses written list of goals as a guide to developing or selecting test items
21a	1-3	Informs students of what will be expected on test
21b	1-3	Gives students samples of questions to be included on the test
22e	1-3	Includes items that require judgment to score
23a	1-3	Gives more weight to items covering materials stressed in unit
22f	1-3	Establishes standards of performance to be met before progressing to next unit
24	1-4	Establishes explicit criteria to judge responses to nonobjective items
23c	1-3	Provides written comments on students' test papers
25a	1-3	Informs students about sections of tests on which they did poorly
25b	1-3	Informs students about sections of test on which they did well
27a	1-3	Uses extra class periods to reteach material that test results showed was misunderstood
27b	1-3 *	Moves on to next unit, regardless of test results, to keep on schedule
28	0-2	Frequently uses corrective instruction when students do poorly on test
29	0-2	Frequently uses enriching practices for students who do very well on test
	12-41	(Range of total points possible)

* The less frequently teachers reported using this practice, the higher was their score on the item.

Table 2-8

Items from the Student Questionnaire Used in Constructing the Impact Scale

Item #	Point Value	Items as Presented to Student
27	1-4	In this class, the teacher makes clear the things I should be studying for the test.
41	1-4	In this class, the teacher gives notice about what will be on a test enough in advance for me to prepare for it.
20	1-4	The tests given by the teacher in this class cover what I expect them to cover.
25	1-4	The scores I get on the tests in this class closely reflect what I have learned.
28	1-4	When I miss something on a test in this class, the teacher gives me specific feedback on what I need to study again.
18	1-4	When a student gets a low score on a test in this class, the teacher makes sure he or she gets the help needed to do better.
22	1-4	When a student gets a low score on a test in this class, the teacher gives makeup work.
	7-28	(Range of total points possible)

Table 2-9

Correlations Among Scores from the Three Impact Scales Before Training

N=30

	Observational Measure	Teacher Questionnaire
Teacher Questionnaire	.52**	
Student Questionnaire	-.14	.26

Table 2-10

Correlations Among Scores from the Three Impact Scales After Training

	Teatment Group (N=15)		Control Group (N=15)	
	Observational Measure	Teacher Questionnaire	Observational Measure	Teacher Questionnaire
Teacher Questionnaire	.47*		.29	
Student Questionnaire	.16	-.10	.32	.35

* p < .05
** p < .01

Table 2-11

Correlations Between Scores from Each of the Two Observations Made Before and After Training

Variable	Correlations Between Pretraining Scores	Correlations Between Post-Training Scores
	N = 60 Observations	N = 59 Observations
Unit learning displayed, read to, or shared with students through class handouts	.35*	.32
Test based upon written statements of learning goals	.38*	0
Essay questions or open ended problems comprised 10% or more of total point value of unit test	a	- .08
Teacher scored essay questions guided by rating scales, explicit criteria, exemplary responses of other clear standards of quality	a	a
Teacher provided written feedback directly on student test papers	.63**	.49*
When reporting results to class, teacher referred to separate sections of tests	- .28	- .12
Teacher focused review on student misconceptions	.16	- .14
Special work given to class based upon students' test performance	a	a
Special work given to individuals or small groups based upon students' test performance	a	a

* $p < .05$
** $p < .01$

[a] This practice occurred so infrequently that it was not meaningful to calculate a correlation.

The hypotheses entertained in the st''dy, however, focused on specific practices. In order to test these hypotheses, the items on the impact scales that pertained to each hypothesis were identified. These items, grouped by hypotheses, are shown in Table 2-12. Note that hypotheses six and seven have been divided into subhypotheses. This was done because responses to items pertaining to these hypotheses clustered, suggesting that the general practices described by the hypotheses were not unitary, integrated practices, but a set of somewhat separate, independent practices.

Correlations among the items pertaining to individual hypotheses are shown in Appendix F.

Table 2-12

Items from the Impact Scales Pertaining to Each Hypothesis

Hypotheses		Observation	Teacher Questionnaire	Student Questionnaire
Compared to teachers in the control condition, teachers in the treatment condition will:				
1.	Communicate more cleary to students the learning goals they are expected to achieve.	3b	-	
2.	Communicate more clearly to students what, how, and when learning is to be assessed.	-	21a 21b	27 41
3.	Use tests that reflect more closely the learning goals they have established.	3a	22c 23a	25 20
4.	Use a greater variety of test formats.	1f	22e	-
5.	Establish clearer criteria for evaluating student responses to essay questions and other items that require judgment to score.	II.1	24	-
6.	Provide more specific feedback to students on what they have and have not learned.			
6.A	more frequently provide written feedback on test papers	2c	23c	
6.B	more frequently provide information on results from specific sections of a test	I.1	25a 25b	- -
6.C	more frequently provide information on learning deficits and ways of dealing with them	III.2	-	28

Table 2-12 (continued)

Hypotheses		Observation	Teacher Questionnaire	Student Questionnaire
7.	More frequently use test information as a guide to instruction			
7.A	for the class as a whole	III.3	22f 27a 27b	
7.B	for small groups or individuals	III.4	28 29	18 22

CHAPTER THREE

RESULTS

Participants' Evaluation of the Training Program

At the conclusion of each school-based training session, all regular teachers (as distinct from lead teachers) completed a six-item questionnaire designed to assess their reactions to the training. In districts A, C, and D, two regular teachers in each treatment school participated in training and completed the questionnaire. In Districts B and E, four regular teachers in each treatment school took part in training and completed the questionnaire.* Thus, the total number of respondents to the questionnaire was 14.

Here is a summary of teachers' responses to each item on the questionnaire.

Question 1: How clear were the goals for the work sessions?

Mean response: 3.43 on a 4-point scale in which 1 = unclear and 4 = very clear
Range: 2 to 4

The two teachers who circled 2 on the scale wrote the following comments: "As time progressed I felt more comfortable with what the tasks were"; "More advance information about expectations would have been helpful."

Question 2: How well organized were the activities?

Mean response: 3.5 on a 4-point scale in which 1 = disorganized and 4 = very well organized
Range: 3 to 4

The participants praised the work sessions for the preestablished time schedule given to them early on the first day.

*In Districts B and E, building administrators requested that two teachers participate in addition to the two who were originally scheduled to participate. These additional four teachers were not considered part of the research sample. No observations were made in their classrooms. They neither completed the teacher questionnaire nor administered the student questionnaire. This was because the project's budget did not permit additional data collection. However, since these teachers did participate in training, their reactions to the training were elicited.

Question 3a: How helpful was the Handbook?

Mean response: 3.15 on a 4-point scale with 1 = not helpful and 4 = very helpful
Range: 2 to 4

Five of the participants admitted that they had not had enough time to give the handbook a thorough examination. Informally, in group discussions, the participants indicated that they would need time to determine how useful the handbook would be when they were setting goals, constructing tests, and analyzing test information outside the workshop setting.

Questions 3b: How helpful were the supplementary materials (e.g., science test items, studies on mastery learning)?

Mean response: 3.14 on a 4-point scale with 1 = not helpful and 4 = very helpful
Range: 2 to 4

The three participants who wrote comments in response to this question indicated that they found the materials to be helpful, but one teacher concluded, "To implement this material, it would need to be taught as a course."

Question 4: How valuable were the work sessions?

Mean response: 3.77 on a 4-point scale with 1 = not valuable and 4 = very valuable
Range: 3 to 4

Participants rated the value of the sessions quite highly. Comments such as the following were made:

- "I got new ideas on feedback and test evaluation which I can implement in my classroom."
- "We had time to discuss and work together on assessment."
- "The goal writing causes one to reevaluate where one is going and reassess the evaluation process."
- "Would like to have more time or attend another session to work on another unit and develop our materials."

The only negative comments centered around the time chosen for the work sessions (many participants wished sessions could have been held in August or early in the fall) and the short duration of the sessions.

Question 5: What was the most important benefit of the work session for you?

All fourteen participants wrote comments in response to this item. Several reported more than one benefit of the work sessions. Four general benefits were identified:

- Insights from working with peers. One teacher wrote, for example, that the greatest benefit was "doing a critical analysis of cur tests together and working as a team."

- Greater understanding about ways of formulating learning goals and relating tests to goals. One teacher wrote that a key benefit was "dealing with goals for a unit and writing a test for those goals."

- Increased skill in writing test items. A representative comment was that the workshop "gave me valuable insight into writing better test questions and what questions really measure."

- General knowledge about the potential role of tests in the instructional process. For example, one teacher wrote, "I gained a different perspective on the application of tests for learning after unit materials had been covered, e.g., using the test as a teaching aid and a diagnostic tool."

Question 6: How do you think the sessions could have been improved?

By far the most frequently proposed improvement was to allow more time for teachers to work on their goals and unit tests and to exchange their products with colleagues. Other proposals included: "Provide more evaluation of tests constructed by teachers," "Consolidate some of the handouts," and "Discuss word-processing techniques for test development."

Level of Program Implementation

In this section information is provided on the degree to which the design of the inservice program was carried out. There were three phases of the program. The first consisted of training for lead teachers and building administrators. The second consisted of school-based training for regular teachers. The third involved interaction among participating teachers and between teachers and building administrators regarding the use of the practices introduced through training.

With respect to phase 1, the design was carried out as planned, with the exception of two small problems associated with Districts B and D. In District B, a few days before training of lead teachers was to take place, the teacher who originally agreed to serve as lead teacher declined this position. Although a capable replacement was quickly found, the new lead teacher did not receive as complete an orientation to the project as the other lead teachers in the study. Nor were the background materials presented to him in sufficient time for them to do any good. Also, this school's principal, who in the spring of 1984 was highly enthusiastic about the project and indicated an interest in attending the inservice sessions for lead teachers, decided to take a sabbatical leave during the 1984-85 school year. Responsibility for carrying out the role the principal originally planned to play was delegated at the last minute to an assistant principal. This administrator was very supportive of the project, but he had almost no time to prepare for his role in it.

The other small problem was that the principal from District D who planned on attending the training session had accidentally circled in his calendar the wrong day for the training. He therefore did not attend. Two members of the project staff visited him several days later to discuss the issues that had been dealt with in the training session. The principal was apologetic about missing the training session and was very receptive to the visit.

After the training sessions, project staff met individually with each lead teacher to assist him or her in preparing for the school-based workshops. These conferences generally lasted between 50 and 90 minutes. Although none of the lead teachers felt totally "on top" of the content of the training program, each felt sufficiently comfortable with it to facilitate the workshops with his or her colleagues.

Phase 2 of the program, which dealt with the school-based workshops, also went according to plan. In each treatment school, the lead teachers carried out their roles successfully, as evidenced by the positive reactions to the training that participants reported (see the preceding section). Also, a building administrator in each school attended a portion of the training sessions, as had been agreed to at the outset of the project.

Phase 3 of the program pertained to activities following the school-based training. These activities were somewhat loosely defined in the training program under the assumption that teachers and administrators would want to shape follow-up activities according to school priorities and individual needs. Project staff suggested that the lead teachers meet once with their colleagues to review plans for implementing new or refined practices, and at least once to discuss how implementation was progressing. It was also suggested that the lead teacher confer at least once with the building administrator who was participating in the project to review progress and discuss important issues that might have arisen. No set agenda or format for these meetings was prescribed.

Finally, the project staff proposed that lead teachers might wish to consider sharing with their department colleagues at the end of the year what had been learned through the project and its possible implications for the department as a whole.

Table 3-1 shows the number and type of posttraining project-related meetings carried out in each treatment school.

Only in District E did participants appear to discuss thoroughly what was learned through training and what its impact was in the classroom. In fact, administrators in the treatment school in District E, in light of discussions about the utility of the training, asked the lead teacher to conduct another set of workshops on integrating teaching and testing for staff members who had not participated in the original inservice program.

In Districts A and C, discussions tended to focus on the need for computer-assisted test-scoring services rather than on the progress being made in classroom use of the targeted practices. In both of these sites, teachers saw benefits in analyzing and reporting information on students' performance in reference to specific learning goals, but they felt that they could not accomplish this without the aid of test-scoring machines and related software programs that would provide goal-based test reports.

Table 3-1

Number and Type of Post-Training, Project-Related Meetings Carried Out in Each Treatment School

District	Number of Meetings among Teachers	Number of Meetings between Teacher(s) and Administrators	End-of-Year Sharing
A	2	1	Yes
B	0	0	No
C	1	2	No
D	1	1	No
E	2	2	Yes

Teachers conveyed their interests and needs to building administrators, who in turn talked to representatives from test-scoring services about the feasibility of obtaining the resources teachers wanted. These discussions are continuing at the time this report is being prepared. While the search for technical resources was an encouraging by-product of the inservice program, it seems in these districts to have displaced to a large extent any other kind of follow-up to the inservice training.

The quality of interaction among each school's project participants regarding the use of targeted practices in the classroom is indicated in Table 3-2. Using evidence obtained through discussions with administrators and lead teachers, two staff members independently assigned ratings. There was 100% agreement in the ratings assigned by the two staff members.

Effects of the Program

Effects of the program were investigated at two levels. The first examined effects on teachers' test-related practices in general. The second examined effects on each of the specific practices identified in the hypotheses for the study.

General Effects

Teachers' scores on each of the three impact scales (derived from the observational measure and the teacher and student questionnaires)* were used to assess the effects of the intervention on teachers' test-related practices as a whole.

Means and standard deviations on each of the three measures are shown in Table 3-3. The table indicates that the mean score on the observational measure was very low before training in both the treatment and control groups. After training the mean increased slightly in the treatment group and declined slightly in the control group.

The pretraining mean score on the teacher questionnaire was about in the middle of the scale in both experimental groups. After training, the mean increased slightly

* For ease of reading, we will refer to the individual impact scales as the observational measure, the teacher questionnaire, and the student questionnaire, even though, as discussed earlier, the scales actually consisted of only a portion of the items included on the original measures.

Table 3-2

Quality of Post-Training Interaction among Project Participants in Each School in the Treatment Condition

0 = No significant interaction around use of practices in classroom

1 = Brief exchange of perceptions and comments about use of practices in classroom

2 = Considerable discussion and planning related to use of practices in classroom

District	Interaction among Teachers	Interaction between Teachers and Administrators
A	1	1
B	0	0
C	1	1
D	1	1
E	2	2

Table 3-3

Means and Standard Deviations on Each Impact Measure

		Treatment Group (N=15)		Control Group (N=15)	
		Pre	Post	Pre	Post
Observational Measure (scale 0-18)	Mean	2.6	3.4	3.2	3.1
	S.D.	1.5	1.8	1.7	1.6
Teacher Questionnaire (scale 12-41)	Mean	24.8	25.4	26.1	24.3
	S.D.	4.3	4.9	3.9	2.5
Student Questionnaire (scale 7-28)	Mean	17.9	17.4	19.4	19.6
	S.D.	1.9	1.8	2.5	2.0

in the treatment group and decreased slightly in the control group.

With respect to the student questionnaire, pretraining mean scores were near the middle of the scale in the treatment group and somewhat higher in the control group. After training, the mean declined slightly in the treatment group and increased slightly in the control group.

To test the significance of the differences between the scores in the two experimental groups, analysis of covariance procedures were used on each measure. Summaries of the analysis of covariance for each measure are shown in Tables 3-4, 3-5, and 3-6.

Table 3-4 indicates that there were no effects on the scores from the observational measure that could be attributed to the experimental treatment. The table also shows that the assumptions of homogeneity of variances and homogeneity of regression (the test for covariate by treatment interaction) were met.

Table 3-5 indicates that the treatment effects on the scores from the teacher questionnaire approached, but did not meet, .05 significance level. Also, the results show that the assumption of homogeneity of variance on the pre-training measure was met. However, statistically significant differences were found between the variances of the post-training measure. This violation of the assumption of homogeneity of variance suggests that the ANCOVA procedure may not have been appropriate. This caution must be tempered, though, by the fact that this violation occurred with the post-test measure and that the groups were of equal sample size.

Given the uncertainty of the appropriateness of the analysis of covariance for the teacher questionnaire, a secondary procedure was performed. A comparison of the mean gain scores of the treatment and control groups was calculated using a simple t-test. The results of this test are also reported on Table 3-5. The results show strong treatment effects. The mean gain of the treatment group was signficantly higher than the mean gain of the control group (p<.01). However, the significant effects

Table 3-4

Analysis of Covariance Summary for the Observational Measure

N=30

Source	df	MS	F	p
Treatment	1	2.32	.99	.25
Residual	1	2.35		
Homogeneity of variances (pre)			1.284	NS
Homogeneity of variances (post)			1.2656	NS
Covariate by treatment interaction			.10	NS

Adjusted post-training means:

Treatment group	3.55
Control group	2.98

Table 3-5

Analysis of Covariance Summary for the Teacher Questionnaire

N=30

Source	df	MS	F	p
Treatment	1	29.054	3.9838	.057
Residual	27	7.378		
Homogeneity of variances (pre)			1.216	NS
Homogeneity of variances (post)			3.842	.001
Covariate by treatment interaction			.592	NS

Adjusted post-training means:	
Treatment group	25.83
Control group	23.84

Comparison between mean gain scores for the treatment and control groups (df=28)
t=3.08, p<.01

Table 3-6

Analysis of Covariance Summary for the Student Questionnaire

N=30

Source	df	MS	F	p
Treatment	1	7.698	4.562	.042
Residual	27	1.687		
Homogeneity of variances (pre)			1.73	.05
Homogeneity of variances (post)			1.23	NS
Covariate by treatment interaction			1.82	NS

Adjusted means:

Treatment group	17.949
Control group	19.034

Comparison between mean gain scores for the treatment and control groups (df=28) t=1.236, p=.222.

appeared to be due, for the most part, to a decrease in scores on the part of the control group (a mean decline of 1.867) rather than to a major gain on the part of the treatment group.

Table 3-6 reports the results of the analysis of covariance on the student questionnaire data. The difference between the treatment and control groups was significant (p=.042). However, the difference was in the opposite direction from what one would have expected. Students in the classes taught by members of the control group scored higher on the post-training measure than did students in the classes taught by members of the treatment group.

But results from this analysis are of questionable validity. Note that the assumptions underlying the analysis were not satisfied; the pretest variances differed significantly from one another (p=.05).

Because the assumption of homogeneity of variances was violated, a supplemental t-test analysis comparing the gain scores of both groups was conducted. The results of this test indicate that there were no significant differences in the gains of the treatment and control group (p=.22). The gain score analysis contradicts the results of the analysis of covariance and strongly suggests that the differences in sample variance on the pretest may be the cause of the significant findings on the student questionnaire.

Effects by Hypothesis

As discussed in chapter two, a measure for each hypothesis was created by selecting pertinent items from the impact scales. Items from at least two of the three impact scales were selected to assess each hypothesis, except in the case of hypothesis one, which was assessed by only one item.

In order to test individual hypotheses, raw scores on the items pertaining to each hypothesis were transformed into linear T scores. The reason for this was that the scales used on each of the measures differed considerably and therefore, if directly combined, the items would not contribute equally to the resulting sum. For example, items on the observational measure had a scale of 0 to 2, whereas items on the student questionnaire had a scale of 1-4. If combined directly, the scores of the student questionnaire would outweigh those on the observation measure. T scores, which have a mean of 50 and a standard deviation of 10, represented a common reference point across each of the scales.

Tables 3-7 and 3-8 show the mean gains made by each experimental group from pretraining (fall) to post-training (spring) with respect to each hypothesis and subhypothesis. To calculate the gain scores the pretraining T score on each item for each teacher was subtracted from the post-training score. Gain scores for all teachers in each experimental condition were then averaged for each hypothesis and subhypothesis.

To test the significance of the differences in gains made by each experimental group, analysis of variance procedures were used. Results from this analysis are summarized in Table 3-9. The table indicates that hypotheses one through six were not supported. Hypothesis seven, however, was supported. This is largely because of the effects obtained for subhypothesis 7A, which concerned teacher's use of test information to guide instruction for the class as a whole. Hypothesis 7 B, concerning teacher's use of test information to guide instruction for individuals or small groups, was not supported.

Although it is encouraging that hypothesis 7 A was confirmed, further analysis of data from the observational measure and the teacher questionnaire indicated that no more than only three of the fifteen teachers in the treatment group made greater use of test information after training than before. This small increase in the number of treatment teachers using test information to guide instruction was accomplished by a small decline in the number of control group teachers using this practice. Because there were only a few occurrences of this practice, these small changes were exaggerated

Table 3-7

Contrast Between Gains Made from Fall to Spring by the Treatment and Control Groups on Measures Pertaining to Each Hypothesis

Hypotheses (Abbreviated)	Treatment Group (N=15)		Control Group (N=15)	
	Mean Gain[a]	(S.D.)	Mean Gain[a]	(S.D.)
1. Communicate learning goals to s'udents	+9.5	(12.9)	+6.2	(14.1)
2. Communicate what, how, and when learning will be assessed	-1.6	(10.3)	+0.7	(9.8)
3. Match tests to goals	+0.5	(11.3)	+1.8	(9.4)
4. Use a variety of test formats	+2.5	(13.2)	+1.2	(11.2)
5. Establish clear criteria for scoring essays and open-ended problem solving items	-3.3	(14.8)	-4.6	(12.7)
6. Provide specific feedback to students on their learning progress	-0.9	(10.4)	-1.7	(10.7)
7. Use test information to guide instruction	+1.0	(9.2)	-3.7	(13.5)

[a] Gains were calculated using T scores

Table 3-8

Contrast Between Gains Made by the Treatment and Control Groups on Measures Pertaining to Each Subhypothesis

Subhypotheses (Abbreviated)	Treatment Group (N=15)		Control Group (N=15)	
	Mean Gain[a]	(S.D.)	Mean Gain[a]	(S.D.)
6A. Provide written feedback on test papers	+2.6	(11.7)	+1.5	(10.9)
6B. Provide information on results from specific sections of a test	- .2	(9.1)	-2.8	(11.2)
6C. Provide information on learning deficits and ways of dealing with them	-4.8	(10.3)	-2.9	(9.4)
7A. Use test results as guide to class instruction	+1.2	(10.1)	-4.2	(12.1)
7B. Use test results as a guide to group or individual instruction	- .1	(8.4)	-3.0	(14.7)

[a] Gains were calculated using T scores

Table 3-9

Summary of Analysis of Variance of Measures Pertaining to Each Hypothesis and Subhypothesis

Hypotheses	df	MS	F	p
1	1	58.4	.318	NS
2	1	154.7	1.524	NS
3	1	21.4	.205	NS
4	1	27.9	.186	NS
5	1	26.4	.139	NS
6	1	25.3	.228	NS
7	1	720.9	5.275	.03
Subhypotheses				
6A	1	14.4	.113	NS
6B	1	149.5	1.436	NS
6C	1	52.8	.541	NS
7A	1	801.8	6.450	.02
7B	1	318.0	2.207	NS

by the transformation of the data into T scores. So, although the treatment effect related to hypothesis 7A was statistically significant, the change appears to have been quite small on a practical level.

Supplemental Analysis

Pre- and post-training scores on each item on each impact scale are shown in Appendices G, H, and I. This item by item breakout provides information on individual behaviors or perceptions. It shows, for example, that students on the whole had much more positive perceptions about teacher's communication of test expectations and about the trustworthiness of test scores than they had about the feedback and assistance that teachers provided following testing. The tables also provide confirming evidence that participating teachers rarely used essay questions or open-ended, problem-solving items, or provided corrective instruction to individuals or small groups. A more extensive exploration of relationships among items on the teacher and student questionnaires is contained in the companion study to this report (Duckworth, Fielding & Shaughnessy, 1986).

CHAPTER FOUR

DISCUSSION

This chapter is organized around three questions: (1) Why did the intervention produce such small effects? (2) How might the staff development program be strengthened? and (3) What are the implications of the present study for further research and development?

Why Such Small Effects?

As reported in the last chapter, the intervention seems to have had very modest effects. Six of the seven hypotheses of the study were not supported, and support for the seventh was small. To the extent that change did take place, it seemed to be confined to a relatively small number of teachers and a relatively small number of practices. Changes resulting from the inservice program were neither as widespread nor as systematic as hoped for.

In this section we discuss limitations in the intervention and the conditions under which it was carried out that might explain why the effects of the staff development program were so small. These include limitations in:

1. the training program
2. the onsite support following training
3. school norms, policies, and incentive systems
4. resources available to teachers
5. teachers' working conditions.

Training Program

The training program, although given high rating by teachers, appears in retrospect to have been weak in three respects. One weakness, frequently mentioned by participants, was that the program was too brief, given its complex objectives. Teachers by and large needed more time than the program provided to translate what they were learning into useful and high quality teaching and testing materials. Teachers also needed more time to exchange work with colleagues and refine products in light of peer review.

Perhaps another weakness in the training program was that lead teachers, from whom so much was expected, received little more training than regular teachers. Yet evidence suggests that teachers need special training and support if they are to make the transition from teaching children or adolescents to teaching peers. In a study of teachers who had assumed the role cf instructional change agents in their school districts, Beaton (1985) described the complex set of skills that classroom teachers need to learn in order to become effective teacher trainers. Beaton observed, for example, that peer teachers must be adept in building and maintaining personal rapport with colleagues, while at the same time challenging them to stretch and grow intellectually and to refine and extend current practice. Beaton indicated that skill in peer teaching generally develops over a period of several years. In retrospect, it seems unrealistic to have expected the lead teachers in the present study to possess fully, after only two days of training and one follow-up planning session, the knowledge and skills needed to foster substantial change on the part of their peers.

Finally, the timing of the training was less than ideal. Lead teachers received training in late November; regular teachers were trained between December and the end of February. By these dates, teaching and testing practices for particular classes were relatively well established. Several participants indicated that the training would have been more beneficial in August when teachers were more open to new ideas and procedures.

Follow-Up Support

There is considerable evidence that teachers need high levels of feedback and technical assistance when attempting to implement a new and complex set of classroom practices (Showers, 1984). The design of the staff development program carried out in this study called for lead teachers and building administrators to provide some assistance and feedback to regular teachers, but the specific quality or nature of this support was not specified. Perhaps as a result of the vagueness of this expectation, and of the limited training lead teachers received generally, little technical assistance or feedback to regular teachers was furnished.

School Norms, Policies and Incentive Systems

Although the schools participating in this study appeared to be well managed, there was very little in the environment of the schools that encouraged teachers to examine or question their current test-related practices or that demonstrated schoolwide commitment to strengthening the ties between instruction and assessment. None of the schools had established testing programs that were directly related to the curricula teachers were teaching. None of the schools had articulated policies on the uses to be made of test information in the instructional process. No school had established a mastery-based grading policy, under which grades were to reflect student performance in relation to pre-established standards of proficiency or excellence. No school had clear, specific policies supporting the practices dealt with in the intervention.

The lack of supporting school policies and norms set limits on the potential impact of the intervention. For example, teachers receiving training were expected to formulate explicit performance standards for a class and hold students accountable to them. But this practice was clearly one that had implications for the organzation and culture of a school as a whole. Teachers are not likely to insist that students meet high and explicitly formulated performance standards when their school and district seem to have no clear commitment to applying such standards. Powell and his colleagues (1985) pointed out that the typical high school in this county is "profoundly neutral about mastery. No one opposes it, but few require or expect it" (p. 61). Powell and his colleagues also noted that student "failure comes from not attending or not behaving. Performance is remarkably irrelevant" (p. 59). In the absence of any clear, schoolwide consensus about the achievement standards to which students must be held accountable, a teacher inservice program emphasizing the importance of mastery standards stands little chance of having a major impact.

The schools also lacked an incentive system through which teachers might be rewarded for using the practices introducedi in the workshops. The only motivation for adopting a goal-based model of integrating teaching and testing was a teacher's belief or feeling that the model was a good one and that it would improve the quality of instruction in his or her classes. But while teachers on the whole recognized the value of the model, the promise that it would enhance student learning was a somewhat abstract motivator, particularly in light of the additional time and effort that use of the model required. If a teacher decided to integrate teaching and testing in an exemplary manner, he or she would be no more likely to receive a promotion, salary increase, or ot er recognition of merit than a teacher who continued to use traditional practices.

Resources

One of the primary concerns of participants in the study was that they lacked the resources needed to carry out effectively some of the practices dealt with in the workshops. Specifically, teachers indicated that they would have benefited from desktop test scoring machines, particularly if these machines interfaced with a microcomputer programmed to report test information on a goal-by-goal basis or in reference to preestablished mastery standards. Some of the teachers also expressed interest in developing goal-referenced, test-item banks in science that users could file in a computer and draw upon according to their particular testing needs. Administrators in the participating schools were receptive to these ideas and in at least two of the schools worked with teachers to find out more about what test-related technology was available and what might possibly by obtained, given district and school priorities and budgets. Teachers in at least one school also made inquiries about a project, sponsored by the Northwest Evaluation Association, designed to create a large test-item pool in science, grades 1 to 12, for use by teachers and administrators.

Teachers also lacked the kind of instructional resources they believed were needed to accommodate effectively both high- and low-achieving students. Many teachers commented that unless special instructional materials geared to students at different levels of learning were made available, they were not likely to group students for corrective or enriching instruction. Developing plans and materials for a class as a whole was difficult enough, it was indicated; to make two or three lessons for a class to meet the special needs of slow- and fast-learning students was considered impractical and overly time consuming.

Teachers' Work Load

Much has been written about the large number of students that high school teachers must work with each day; the variety of subjects or courses they must teach; the limited amount of time they have for planning, reflection, or interacting with colleagues; and the numerous noninstructional duties, such as monitoring corridors and lunchrooms, that they often must carry out (Sizer, 1984; Darling-Hammond, 1984). The practices recommended in the staff development program required teachers to invest more time and effort in the teaching-testing process than they ordinarily did. But there was little slack time in teachers' schedules. Most teachers reported that there simply wasn't sufficient time to do all that was called for in the training program.

Ways to Strengthen the Program

The staff development program we offered might be strengthened in four ways: (1) anchoring the program to schoolwide policies and commitments; (2) providing more extensive training for lead teachers and administrators who will be responsible for implementing the program; (3) providing more time for training sessions; and (4) carefully structuring support systems for teachers participating in the program.

Anchoring the program to Schoolwide Priorities and Commitments

The present project was a research study. It had little connection to the day-to-day operation of the school programs or with the policies of participating school systems. To be sure, participating schools and teachers chose to take part in the program because at least some school personnel viewed it as important, but school staff had little "ownership" in the project. Participating teachers in the program were free to use or to reject any or all of the practices introduced through training. The message that project staff communicated was: "Here are some research-based principles, guidelines and procedures for improving teaching and testing that you might wish to consider. We will provide illustrations of how these ideas can be applied in classrooms and give you time and some assistance in making your own applications. However, what we offer are not prescriptions but possibilities for you to consider. What you do with these possibilities is totally up to you." Thus, the inservice program was disconnected from school policies, priorities, or commitments. It was an isolated event in the lives of participating teachers and schools.

To increase the effects of the inservice program, a school or district must make a broad commitment to the underlying principles of a goal-based, mastery-learning model of instruction. It is not enough to present the model simply as a set of possibilities for individual teachers to consider. Teachers, administrators, parents, and local boards must be clear aobut the implications of the goal-based, mastery-learning model and reach a common understanding of what should be done in schools and classrooms to implement the model. The kinds of resources provided through the staff development program carried out in the present study can assist school personnel in the implementation process, but they cannot substitute for schoolwide policies and expectations.

It should be noted that the importance of linking the inservice program to broader policies and understandings is discussed in The Planning Guide for Lead Teachers and Administrators, which currently accompanies the training materials on integrating teaching and testing. The guide is available through the Teaching Research Division, Oregon State System of Higher Education. The need for schoolwide communication about the nature and implications of a goal-based, mastery-learning model of instruction is emphasized in this document, as is the need to establish policies and expectations regarding the model's use. The planning guide was completed when the intervention described in this report was nearing completion; thus, it reflects lessons learned form the present study, as well as understandings gained from related work in this area (Fielding and Schalock, 1985b; Schalock et al., 1985).

Providing More Extensive Training for Lead Teachers and Administrators

Most candidates for a lead teacher position probably need at least one year to prepare for this role. The kind of preparation needed is described in The Planning Guide for Lead Teachers and Admininstrators. It goes well beyond the two days of training offered in the present study. It must involve sustained self-study; substantive interaction with knowledgeable experts in the district and, to the extent appropriate, at institutions of higher education; and ongoing efforts to apply new practices in the classroom. In addition, as indicated above, lead teachers must be clear about the role they are to play in their departments and schools, or their preparation will be somewhat aimless. Ideally, at least two lead teachers would be preparing for lead teacher roles at the same time so that a peer support system could develop. As Beaton (1985) noted, successful school- or district-based instructional change agents commonly work in pairs, and depend heavily on one another for both technical and emotional support.

Administrators also need more training than what was furnished them in the current project. Building administrators do not have to be technical experts in integrating teaching and testing, but they must know how to bring about the conditions that foster and support this integration. The responsibiliites that administrators need to assume in a staff development program of the kind carried out in the present study are also described in The Planning Guide for Lead Teachers and Administrators.

Providing More Time for Training

On the basis of findings from the present study, it appears that at least three days of training, rather than two, are needed to deal adequately with the objectives of the school-based workshops for regular teachers. As mentioned earlier, more time is especially needed for teachers to translate ideas into working materials and to exchange these materials with colleagues.

Ideally, teachers would receive training before school opens in the fall, so that they would be able to introduce students to the goal-based model at the beginning of a course rather than after classroom teaching and testing patterns had been established.

Carefully Structuring Support Systems for Teachers

Reports on successful mastery-learning programs (Fitzpatrick, 1985; Little, 1984; Westerberg and Stevick, 1985) indicate the ongoing collegial support and interaction regarding the implementation and short-term effects of the program is absolutely essential to the program's success. After training in a mastery-learning approach to integrating teaching and testing is completed, teachers apparently need to meet regularly throughout the year to review progress in implementing new practices in the classroom and to discuss and resolve implementation issues they encountered. In view of results from the present study, it seems that one cannot expect these meetings to occur spontaneously; building administrators or district staff need to organize and help focus such meetings, providing support and reinforcement on the one hand, while insisting on growth and improvement on the other. The need for structured and sustained collegial support is discussed in The Planning Guide for Lead Teachers and Administrators, referred to earlier.

The proposals discussed above for strengthening the staff development program investigated in the present study are silent about the need for changes in teacher incentive systems and teacher work loads, and about the need for additional resources for instruction and assessment. This is not to imply that these needs are small or inconsequential. They obviously deserve deliberate and serious attention. But few schools can deal simultaneously with each and every need for improvement. The discussion here has focused on those changes that seem most immediately necessary to strengthen the inservice program that was offered. Additional improvement in school conditions and resources would have to be made in the long run to achieve major improvements in instruction and learning.

Directions for Future Research and Development

We would make three recommendations for further research and development; One is that a series of case studies be carried out that focus on the complex set of district, school, and classroom factors involved in implementing a goal-based, mastery-learning approach to instruction in high school. As discussed above, there seems little point in trying to provide training to teachers in this approach to instruction unless district and school policies and procedures are established that support the approach. Instead of conceiving of the staff development program as a "stand alone," one-year training program, it probably needs to be viewed as part of a larger school improvement effort. If several schools could be identified that were willing to mount the kind of extensive improvement effort needed to foster the integration of teaching and testing, case studies might be undertaken to document and analyze the change effort and to assess its effects. What is needed is information on how schools can go about implementing and institutionalizing a goal-based, mastery-learning model over a period of at least three years.

We would also recommend that, in the context of the kind of case study proposed above, data collection on the implementation of key teaching and testing practices be used to assist teachers in analyzing and improving their teaching and testing practices. We felt on numerous occasions that participating teachers would have made more use of designated practices if the observers who collected data from the teachers' classrooms could have shared the data with them. The logic of the experimental design used in this study precluded the sharing of data with teachers while the study was in progress, but we sense that this would have been a powerful way of enhancing teachers' understanding of mastery-learning strategies and their interest and commitment in using them.

Finally, it would be worthwhile to investigate the possibility of developing students' understanding of the concepts associated with a goal-based, mastery-learning model of instruction. In this regard, students might receive instruction in the meaning of such concepts as "learning goal," "measure of goal attainment," and "performance standard." They might also profit from instruction in setting their own goals for learning, or setting goals cooperatively with teachers, as is required in independent-study projects, and in determining how they themselves and others might assess their learning progress. Teaching s udents how to profit from constructive feedback on their work and how to furnish such feedback to their peers might also be a worthwhile target of training. Perhaps one of the reasons why there was such a low correlation in the present study between teachers' and students' perceptions of test-related practices was that teachers and students had not worked together to develop a common understanding of these practices and their importance in the learning process. Perhaps training procedures for students need to be developed to complement and reinforce the training in integrating teaching and testing that teachers receive.

REFERENCES

Austin, G.R. (1979). Exemplary schools and the search for effectiveness. Educational Leadership, 37, 10-14.

Barnes, S. (1985). A study of classroom pupil evaluation: The missing link in teacher education. Journal of Teacher Education, 36(4), 46-49.

Beaton, C.R. (1985). Identifying change agent strategies, skills, and outcomes: The case of district-based staff development specialists. Unpublished doctoral dissertation, University of Oregon.

Block, J.H., & Burns, R.B. (1976). Mastery learning. In L. Schulman (Ed.), Review of research in education (Vol. 4, pp. 3-49). Itasca, IL: F.E. Peacock.

Bloom, B.S. (1976). Human characteristics and school learning. New York: McGraw-Hill.

Burry, J., Catterall, J., Choppin, B., & Dorr-Bremme, D. (1982). Testing in the nation's schools and districts: How much? What kinds? To what ends? At what costs? Report No. 194. Los Angeles, CA: Center for the Study of Evaluation, University of California, Los Angeles.

Carter, K. (1984). Do teachers understand principles for writing tests? Journal of Teacher Education, 35(6), 57-60.

Cohen, M., (1982, February-June). Effective schools: Accumulating research findings. American Education, 13-16.

Darling-Hammond, L. (1984, July) Beyond the commission reports: The coming crisis in teaching. Reed Corporation, Report No. R-3177-RC.

Duckworth, K., Fielding, G.D., & Shaughnessy, J. (1986). The relationship of high school teachers' class testing practices to students' study effort and feeling of efficacy. Eugene, OR: Center for Education Policy and Management, University of Oregon.

Farrar, E., Neufield, B. & Miles, M.B. (1984). Effective schools program in high schools: Social promotion or movement by merit? Phi Delta Kappan, 65, 701-706.

Fielding, G.D., & Schalock, H.D. (1985a). The development of a teacher's handbook and a related staff development program for integrating teaching and testing in high school. Final report, Contract #400-82-0013, The National Institute of Education. Monmouth, OR: Teaching Research Division, Oregon State System of Higher Education.

Fielding, G.D., & Schalock, H.D. (1985b). Promoting the professional development of teachers and administrators. Eugene, OR: ERIC Clearinghouse on Educational Management.

Fitzpatrick, K.A. (1985, April). Group-based mastery learning: A Robin Hood approach to instruction? Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Guskey, T.R. (1985). Implementing mastery learning. Belmont, CA: Wadsworth.

Guskey, T.R. & Gates, S.L. (1985). A synthesis of research on group-based mastery learning programs. Paper presented at the annual meeting of the American Education Research Association, Chicago, IL.

Hord, S., Stiegelbauer, S., & Hall, G. (1984). Principals don't do it alone: Researchers discover second change facilitator active in school improvement efforts. R & DCE Review, 2(3), 1, 2, 5.

Hunter, M. (1976). Improved instruction. El Segundo, CA: TIP Publications.

Lawler, E. (1976). Control systems in organizations. In M. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Natriello, G., & Scott, P. (1981). Secondary school evaluation systems and student disengagement. Paper presented at the annual meeting of the American Educational Research Association.

Porter, A.C. (1983). The role of testing in effective schools. American Education, 19(1), 25-28.

Powell, A., Farrar, E., & Cohen, D. (1985). The shopping mall high school. Boston: Houghton Mifflin.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). Eighteen thousand hours: Secondary schools and their effects on children. Cambridge: Harvard University Press.

Ryan, D.W., & Schmidt, M. (1979). Mastery learning: Theory, research, and implementation. Ontario, Canada. Ministry of Education.

Schalock, H.D., Fielding, G.D., Schalock, M., Erickson, J., & Brott, M. (1985). Integrating teaching and testing with program management. Educational Leadership, 43(2), 55-58.

Showers, B. (1984, April). Peer coaching and its effects on transfer of training. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. ED 186 401.

Sizer, T.R. (1984). Horace's compromise: The dilemma of the American high school today. Boston: Houghton Mifflin.

Stiggins, R., & Bridgeford, N. (1982). Final research report on the role, nature, and quality of classroom performance assessment. Portland, OR: Northwest Regional Educational Laboratory.

Westerberg, T., & Stevick, J. (1985). Mastery learning at the high school level: A prescription for success. Outcomes (A quarterly newsletter of the Network for Outcome Based Schools), 5(1), 24-27.

Appendices referred to in this report can be obtained for $3.50 from the Center for Educational Policy and Management, College of Education, University of Oregon, Eugene, Oregon 97403.