DOCUMENT RESUME

ED 242 718 SP 024 155

AUTHOR Stiggins, Richard J.; Bridgeford, Nancy J.
TITLE The Use of Performance Assessment in the Classroom.
INSTITUTION Northwest Regional Educational Lab., Portland, OR. Center for Performance Assessment.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Jan 84
CONTRACT 400-83-0005
NOTE 45p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Classroom Techniques; *Educational Testing; Elementary Secondary Education; Informal Assessment; Standardized Tests; *Student Evaluation; *Teacher Attitudes; *Teacher Made Tests; Test Reliability

ABSTRACT

A study explored the nature and quality of teacher-developed assessment instruments. Teachers (n=228) from a range of grades, subjects, and school districts described patterns of test use, concerns about assessment, and use of performance assessment by completing an extensive questionnaire. The research was conducted to determine: (1) teachers' skills, attitudes, perceptions, and concerns about day-to-day classroom assessment; (2) the extent to which performance tests (versus other forms of assessment) are used in classrooms; (3) the nature of performance tests; and (4) whether (or how) teachers check and/or attempt to improve the quality of their classroom performance assessments. Results suggested that the foundation and structure of classroom assessment consists primarily of teacher-developed assessments, with performance assessment serving as one of the key tools. Five major issues are analyzed and discussed: (1) the use and importance of performance assessment in the classroom; (2) the stability of results across grades, subjects, and research contexts; (3) teachers' concerns about assessment, particularly with respect to improving test quality and use; (4) specific issues of assessment quality, including potential difficulties in classroom performance assessment procedures; and (5) actions needed to overcome some of the assessment problems. (Author/JD)

# THE USE OF PERFORMANCE ASSESSMENT IN THE CLASSROOM*

Richard J. Stiggins
and
Nancy J. Bridgeford



Center for Performance Assessment
Northwest Regional Educational Laboratory
300 S. W. Sixth Avenue
Portland, Oregon 97204

January, 1984

*This research was conducted under contract #400-83-0005 with the National Institute of Education. Opinions expressed in this publication do not necessarily reflect the position of NIE, and no official endorsement should be inferred.

ABSTRACT

Research on classroom assessment has tended to focus on standardized tests and has paid minimal attention to teacher-developed assessments. As a result, we have a narrow understanding of the classroom assessment environment. This study was designed to broaden that understanding by exploring the nature and quality of teacher-developed assessment. Particular emphasis was given to understanding the role of performance assessment--observation and rating of behavior--in the classroom. Teachers from a range of grades, subjects and school districts described their patterns of test use, concerns about assessment and use of performance assessment by completing an extensive questionnaire. When responses are summarized across teachers, the results suggest that the foundation and structure of classroom assessment consists primarily of teacher-developed assessments, with performance assessment serving as one of the key assessment tools. Teachers are concerned about assessment and know that improvement may be needed, but will need help to affect needed changes. Action plans are suggested for enhancing the quality of teacher-developed tests.

# INTRODUCTION

Teachers use many assessment methods to track student growth and development. They devise many of those assessments themselves. Some are paper and pencil measures. Others are based on behavioral observations. This research explores teachers' observation and rating of student behavior and products as these measures relate to the larger context of day-to-day classroom assessment. Specifically, the research was conducted to determine: (a) teachers' skills, attitudes, perceptions and concerns about day-to-day classroom assessment; (b) the extent to which performance tests (behavioral observations and ratings)--versus other forms of assessment--are used in classrooms; (c) the nature of classroom performance tests (uses, exercises, responses and performance rating procedures); and (d) whether (or how) teachers check and/or attempt to improve the quality of their classroom performance assessments.

To date, the measurement community has tended to limit its study of testing in the schools to the role of large-scale standardized testing programs. Far less attention has been given to the nature or quality of teacher-developed classroom assessments. And almost no attention has been given to the nature or quality of observational assessment methods like performance assessment. To illustrate, nearly all major recent studies of teachers' testing practices and attitudes have focused on the role of standardized tests in the education process (Goslin, 1967; Lortie, 1975; Airasian, et al., 1977; Stetz and Beck, 1979; Rudman, et al., 1980; Salmon-Cox, 1981; Sproul and Zubrow, 1981; and Kellaghan, Madaus and Airasian, 1982). Further, a recent special issue of the

Journal of Educational Measurement (Burstein, 1983) on the state of the art in linking testing and instruction is introduced as follows:

> Linking testing and instruction is a fundamental and enduring concern in educational practice...Fundamental questions about how well achievement test items reflect both student knowledge and the content of instruction are clearly at the heart of the matter... [yet] The contributors [to this special issue] were asked to limit their conception of achievement testing to include standardized achievement tests, curriculum embedded or locally developed domain-referenced and proficiency tests, and state assessments. Thus, teacher made tests...were systematically excluded. (p. 99, emphasis added)

Thus, this "state of the art" review linking testing and instruction was constrained to the kind of test information obtained from instruments developed outside the classroom--measures providing only a portion of the data teachers use to integrate testing and instruction.

This emphasis on large-scale and standardized tests on the part of measurement researchers may result from the strong tradition of scientific inquiry in educational research and psychometric models in educational measurement (Coffman, 1983; Calfee and Drum, 1976). These emphases lead to admonitions in our measurement textbooks that teachers should strive to gather "hard data" on student achievement by relying on using only objective tests. Yet, several researchers conclude from their studies of testing in the schools that teachers purposefully go beyond test scores and are intent on using observation-based modes of assessment to acquire information for decision making. For example, in a national study of classroom assessment, Herman and Dorr-Bremme (1982) report, "nearly every survey respondent reported that 'my own observations and students' classwork' was a crucial or important source of information."

In another study, Salmon-Cox (1981) concludes, "overwhelmingly, we found that teachers, when talking of how they assess their students, most frequently mention 'observation'. Clearly this favored teacher technique is quite different from the kind of information provided by standardized tests."

And Kellaghan, Madaus and Airasian (1982) point out "standardized test information appears to represent an auxiliary or secondary criterion in [instructional] judgment, since teachers were nearly unanimous in stating that the most commonly reported grouping criteria were the teachers' own observations and tests."

In fact, the Herman and Dorr-Bremme study, along with that of Yeh (1978), are among the few investigations of testing in the schools to go beyond the role of standardized tests and focus on teacher-developed tests. Their national survey results suggest that, depending on grade level, a third to three-quarters of tests used in the classroom are teacher developed. To understand the implications of the omission of teacher-developed tests from prominent research on measurement practice, we must consider the differences between the views of the test specialist concerned with scientific measurement and the measurement needs of the classroom teacher. Coffman (1983) provides a concise analysis of these differences by referring to earlier comments by Scates (1943):

> Scates pointed out that the scientist is interested in truth leading to broad generalizations while the teacher seeks information of direct practical value; the scientist is interested in elements whereas the teacher is interested in functioning organisms; the measurement specialist cannot measure continuously, but the teacher needs to and must measure continuously; the scientist measures traits uniform throughout their range, but the

> teacher measures growth in stages; and the measurement specialist generally measures formal abilities by cross-section power tests, but the teacher must be concerned with behavioral dynamics in life situations. To the extent that Scates' analysis is sound, it is not surprising that there is little systematic study of teachers' testing practices reported in the literature written primarily by researchers and test specialists.

If measurement researchers continue to emphasize only those tests that serve large-scale assessment purposes, we may fail to serve teachers' primary measurement needs. Measurement training that relies on traditional objective tests does not meet all of the day-to-day assessment needs of teachers. It disregards the full range of measurement options available to teachers and, most important, it fails to help teachers produce data needed to address the day-to-day decisions they face. The research reported here is designed to broaden our understanding of teachers' day-to-day assessment needs.

## The Types of Classroom Assessment Explored

One goal of this research was to determine the role and relative importance of four types of measurement in the classroom: the teachers' own objective tests, published tests, structured performance assessments and spontaneous performance assessments.

The teachers' own objective tests were defined to include those multiple choice, true/false, matching and short answer fill-in tests teachers design for use on a day-to-day basis in their classrooms. Published tests were defined to include both standardized objective achievement tests and objective tests supplied as part of published text materials.

Performance assessment, as defined for the purpose of this research, is testing to be sure, but not in the traditional sense of objective test. Rather, performance assessment calls for the observation and rating of student behavior, and necessitates that students actually demonstrate proficiency (Stiggins, in press).

Performance tests have several important characteristics: First, students are called upon to apply the skills and knowledge they have learned. Second, performance assessment involves completion of a specified task (or tasks) in the context of real or simulated assessment exercises. Third, the assessment task or product completed by the examinee is observed and rated with respect to specified criteria, in accordance with specified procedures.

In the research reported here, we make an important distinction with respect to performance assessment. We distinguish between structured and spontaneous performance assessments. The former is planned and systematically designed to include prespecified purposes, exercises, observations and scoring procedures. The latter arises spontaneously from the naturally-occurring classroom environment and leads the teacher to a judgment about an individual student's level of development.

In this paper, we summarize results from a large-scale survey of teachers' uses of these various testing methods, their concerns about assessment and the specific characteristics of their performance assessments.

## RESEARCH METHODOLOGY

The study was designed to probe assessment practices in a stratified sample of teachers selected from eight districts across the country, varying in size and geographic location. Five districts were urban, three suburban; three were in the East, two in the Northwest, and three in the West. Each district was to recruit 48 volunteer teachers to complete a comprehensive questionnaire on classroom assessment. Twelve teachers were to be recruited from each of four grades (2, 5, 8 and 11). Of those 12 teachers at each grade level, three were to describe their assessment methods in writing, three in speaking, three in science and three in math. Thus, each respondent described assessment methods in only one subject, and at only one grade level.

All districts responded with completed surveys; however, the number of completed forms differed substantially across districts. A total of 228 completed questionnaires were received. The respondents were distributed almost equally across districts, grades and subjects.

Although 228 responses represented less than our desired sample of 384, the group was sufficiently large to proceed with the analysis. In analyzing and subsequent interpreting of data, however, we proceeded with caution for two reasons. First, the sample size precluded an analysis of teachers by subject area within each grade level--8th grade science teachers, for example. Analyses of the responses were limited to grade, subject and district totals. Second, generalizations beyond the volunteer sample were not attempted.

## The Questionnaire

Questionnaire Design. The questionnaire was designed in several steps. First, questions were devised to tap various levels of concern about and use of the basic types of classroom assessment. The initial version of the questionnaire served as the basis for structured interviews with teachers, during which the question underwent extensive revision (Stiggins & Bridgeford, 1982). It was then reviewed and critiqued by numerous teachers, educational researchers, and editors through a long series of revisions and refinements. As a final step, the questionnaire was field tested with 30 teachers from several grades and subjects.

To ensure that teachers understood the meaning of each type of assessment covered in the questionnaire, they were provided with concise definitions of teacher-made objective tests, published tests, structured performance tests and spontaneous performance assessments at the beginning of the questionnaire. In each case, the teacher was asked to supply an example of each kind of test from his or her experience. If the example revealed that the teacher did not understand the definitions of and distinctions between assessment types, that teacher's responses were not included in the analysis. A small number of booklets from each district (usually 2 or 3) were eliminated for this reason.

Levels of Use. One major set of questions probed teachers' use of four specific assessment options. Teachers were asked to describe the importance of different test options as a function of their specific reasons for testing; that is, for diagnosis, grouping, grading,

evaluating instruction, and reporting achievement results. Respondents were given these instructions:

> Describe the relative importance of each type of assessment by indicating the weight you give to each in achieving your various classroom assessment purposes. Each question below identifies a specific instructional purpose. If a certain type of assessment carries no weight in achieving a given purpose, you should enter 0% next to it. On the other hand, if you rely completely on one type of assessment for a specific purpose, you should enter 100% next to that type. As another example, a response of 25% to each of the four indicates equal weight to each in achieving that purpose. Percentages for each purpose should total 100.

Second, to determine the extent to which each assessment option was used, we used an adaptation of a scaling system developed by the University of Texas Research and Development Center in Teacher Education (Hall, et al., 1979) to pinpoint teachers' levels of use of the four alternative assessment methods as indicated in the following scale:

> NONUSE: No action is being currently taken or anticipated with respect to this type of assessment.
>
> ANTICIPATED USE: The user has decided to start using this type of assessment, but has not yet acted upon that decision.
>
> PREPARATION TO USE: The user is preparing to use (studying, taking action to begin using) this type of assessment but is not yet doing so.
>
> EFFORTFUL USE: The user is using that test type, but that use is labored, requiring much effort.
>
> COMFORTABLE USE: The user is using this type of assessment with ease.
>
> REFINING USE: The user is making changes in assessment procedures to increase outcomes, and is working alone on this.
>
> COLLABORATION IN USING: The user is making deliberate efforts to coordinate with others in developing and using this type of assessment.

Scaling on the teachers' level of use was accomplished by having the respondent answer a branching series of questions about their use of each test type.

Types of Concern. We also investigated teachers' concerns about each individual type of test, by adapting the "levels of concern model" developed at the University of Texas Research and Development Center (Hall et al., 1977). This model helps uncover teachers' perceptions of their own assessment needs by asking teachers to identify their primary concern (e.g., lack of information, management issues) about each type of classroom assessment. Possible concerns about teacher-made objective tests, published tests, and performance assessment (structured and spontaneous) as listed below. Each teacher was asked to identify his or her primary concern by selecting from among these statements.

| Teachers concerned about: | Select: |
|---|---|
| Lack of information | I am concerned about my lack of information about developing and using (my own objective paper and pencil tests.) |
| Competence | I am concerned about my level of training, skill and experience in developing and using (my own objective paper and pencil tests.) |
| Time management issues | I am concerned about the amount of time required to manage the development and use of tests. |
| Consequences of use | I am concerned about how my students react when I administer (my own objective tests.) |
| Collaboration in using | I am concerned about establishing working relationships with other teachers to develop and use (objective tests.) |

| | |
|---|---|
| Test improvement | I am concerned about making such tests better and using them more effectively. |

Teachers who had no primary concern were asked to leave the item blank

Teachers' concerns indicate the type of information about testing that is likely to be of greatest interest and use to teachers at any given point in time. For example, if a teacher is concerned about the adequacy of her/his training and skill in assessment, that teacher is unlikely to be interested in strategies for working with other teachers to improve testing. Rather, the competence concern must be satisfactorily addressed first. To assist us in interpreting concerns more accurately, teachers were also requested to cite the specific reason(s) why the response they selected was primary for them.

Performance Assessment. The remaining questionnaire items focused specifically on structured performance assessment. These questions were asked in two forms. First, teachers were asked to give an example of a structured performance test used previously. They were then asked to further describe that example by answering a series of questions about its development, administration, scoring, use and quality. These initial sample questions were designed to ensure that teachers understood the characteristics of performance tests as distinct from other teacher-developed tests. After describing the example, teachers were then asked to answer a parallel set of questions about their general use of structured performance tests. These latter questions (listed in Table 4) provided the specific information which was analyzed in order to understand teachers' use of performance assessments in each subject area and grade level.

## RESULTS

Results are summarized in several parts. First, we report teachers' patterns of test use in terms of the levels of test use scale and the relative weight teachers assigned to different test types for different purposes. The analysis then turns to teachers' concerns about assessment. Concerns of respondents are summarized in terms of (1) the types of concern, and (2) teachers' stated reasons for those concerns. The third part of the analysis addresses teachers' use of structured performance assessments, describing test characteristics and quality control procedures. In all three cases, data are explored across test type (teacher-made objective, published, structured performance assessment, and spontaneous performance assessment), grade level (2, 5, 8, and 11), and subject area (writing, speaking, science, and math).

The overall goal of the analysis is to describe the classroom assessment practices--use, preferences, attitudes, and role of performance assessment--of these 228 volunteer teachers. Since these teachers may not be representative of the general teacher population and since the practices described reflect what teachers say they do--not necessarily what they actually do--inferences about the testing practices of all teachers are not justified.

In all cases, we have attempted to select and discuss the largest, most notable patterns of difference in teachers' responses as they varied across test type, grade and subject. Due to the exploratory nature of the study, limitations in the characteristics of the sample of

respondents, and number and complexity of the questions asked, questions of the probability of occurrence of particular differences were not addressed via statistical analysis.

Patterns of Test Use

Levels of Use. Table 1 reports the percentage of respondents at each category on the level of use scale.

Looking first at teacher-made objective tests, about half of these teachers report comfortable use. This holds across grades and subjects. The other half of the teachers vary in level of use. For instance, use of teacher-made objective tests tends to increase steadily as grade increases (i.e., nonuse percent declines); but teachers may struggle some to increase use of this type of test as indicated by the increase in the effortful use category. Further, math and science teachers tend to use their own objective tests slightly more than writing and speaking teachers.

Note also that (a) about 20% of respondents claim that they do not use their own objective tests, (b) few teachers anticipate use of this test type, (c) few are preparing for future use, and (d) collaboration in use of teacher-developed objective tests is very low. Points b, c and d remain constant for all test types, grades and subjects.

Regarding published tests, again, · rly half report that they use these tests with relative ease, with most of the others reporting that they do not use them at all. There appears to be slightly more use in early grades and appreciably more use in math relative to other subjects. Here again there is no preparation for change and no collaboration.

The levels of use for performance assessment--structured and spontaneous--differ from the objective tests. Eighty-five percent of these teachers report some use of structured performance tests. Forty-eight percent report comfortable use, with another quarter refining their use of these assessments, and 15% of teachers also report effortful use. Nearly 95% of respondents report use of spontaneous performance assessments, with nearly 80% reporting comfortable use. All of these patterns seem relatively constant across grades and subjects.

Role of Test Type as a Function of Purpose. Patterns of reliance on test types vary slightly as testing purpose changes. Table 2 summarizes the relative importance teachers assigned to the various test types for diagnosing the strengths and weaknesses of individual students, grouping for instruction, assigning grades, evaluating the effectiveness of an instructional treatment and reporting results to parents. Since teachers assigned higher percentages to the methods that contribute most to each decision, these data are hereafter called "reliance percentages" in describing and interpreting the results. The higher the reliance percentage, the more weight given to a type of test for that purpose.

For diagnosis, teacher-developed objective tests are reported to be given most weight, with both types of performance assessment close behind. Published tests play a secondary role. Patterns vary across grades. Teacher-made objective tests appear somewhat more important in later grades, while published tests seem somewhat less so. Structured performance assessment is given more importance in diagnosing in grade 11 than in lower grades, while spontaneous performance assessment is reported to be least important at grade 11. Across school subjects, teacher-made objective tests appear most important for diagnosing in

TABLE 1
LEVEL OF USE BY TEST TYPE,
GRADE AND SUBJECT (in percent of respondents)

| Level | Grade | | | | Subject* | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 5 | 8 | 11 | WR | SP | SC | MA | Sample |
| N | 57 | 58 | 58 | 55 | 58 | 61 | 50 | 59 | 228 |
| **Teacher-made Objective Tests** | | | | | | | | | |
| Nonuse | 32% | 26 | 15 | 9 | 26 | 29 | 12 | 14 | 21 |
| Anticipated use | -- | -- | 2 | -- | 2 | -- | -- | -- | .4 |
| Preparation to use | 2 | -- | -- | 2 | -- | 2 | 2 | -- | 1 |
| Effortful use | 5 | 10 | 15 | 25 | 11 | 21 | 8 | 14 | 14 |
| Comfortable use | 53 | 45 | 53 | 47 | 47 | 40 | 61 | 51 | 49 |
| Refinement | 9 | 19 | 15 | 15 | 14 | 9 | 14 | 20 | 14 |
| Collaboration | -- | -- | 2 | 2 | -- | -- | 2 | 2 | 1 |
| **Published Tests** | | | | | | | | | |
| Nonuse | 30 | 25 | 40 | 44 | 34 | 54 | 34 | 15 | 35 |
| Anticipated use | -- | 4 | 2 | 2 | 2 | -- | 6 | -- | 2 |
| Preparation to use | 4 | -- | 3 | 7 | 5 | 3 | 4 | 2 | 4 |
| Effortful use | 9 | 7 | 7 | 4 | 4 | 3 | 4 | 15 | 7 |
| Comfortable use | 49 | 56 | 38 | 35 | 41 | 38 | 38 | 61 | 45 |
| Refinement | 9 | 9 | 9 | 6 | 14 | 2 | 10 | 7 | 8 |
| Collaboration | -- | -- | 2 | 2 | -- | -- | 4 | -- | 1 |
| **Structured Performance Assessment** | | | | | | | | | |
| Nonuse | 17 | 4 | 14 | 8 | 4 | 13 | 11 | 14 | 10 |
| Anticipated use | -- | 2 | 2 | -- | -- | 2 | -- | 2 | 1 |
| Preparation to use | -- | 4 | -- | -- | 2 | 2 | -- | -- | 1 |
| Effortful use | 11 | 14 | 16 | 17 | 18 | 10 | 23 | 9 | 15 |
| Comfortable use | 57 | 51 | 46 | 40 | 52 | 52 | 36 | 52 | 48 |
| Refinement | 13 | 26 | 21 | 26 | 25 | 17 | 23 | 22 | 22 |
| Collaboration | 2 | -- | 2 | 10 | -- | -- | 6 | 2 | 3 |
| **Spontaneous Performance Assessment** | | | | | | | | | |
| Nonuse | 2 | 2 | -- | 9 | 2 | 3 | 4 | 4 | 3 |
| Anticipated use | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| Preparation to use | -- | -- | -- | 2 | -- | -- | 2 | -- | 1 |
| Effortful use | 2 | 5 | 2 | 4 | 7 | 3 | 2 | -- | 3 |
| Comfortable use | 84 | 83 | 82 | 66 | 77 | 85 | 72 | 79 | 79 |
| Refinement | 11 | 9 | 17 | 17 | 14 | 8 | 17 | 14 | 13 |
| Collaboration | 2 | 2 | -- | 2 | -- | -- | 2 | 4 | 1 |

*WR stands for Writing, SP for Speaking, SC for Science, MA for Mathematics

0861e

science and math, while structured performance assessment is most important in writing assessment, while spontaneous performance assessment is given most weight in speaking diagnosis.

When forming instructional groups, on the average, these teachers give approximately equal weight to all four types of tests. However, examination of grade and subject differences reveals some notable variations. For instance, as grade increases, the importance of published tests and spontaneous performance assessment decreases, while weight given to structured performance assessment and teacher-made objective tests increases. Also, for grouping (as for diagnosing), math and science teachers tend to rely on their own objective tests, while writing teachers give most weight to structured performance assessment and speaking teachers rely most heavily on spontaneous performance assessment.

When assigning grades, teacher-made objective tests stand out as most important, followed by structured performance assessment. Published tests and spontaneous performance assessment play lesser roles. Within this pattern, however, there are clear trends across grades. As grade level increases, the weight given to objective tests and structured performance assessment goes up, while that given to published tests and spontaneous performance assessment goes down. Across school subjects, once again, math and science teachers give most credence to their own objective tests, while writing tests rely most on structured performance tests.

In order to evaluate the effectiveness of an instructional treatment, these teachers tend to use their own objective tests, followed by structured and/or spontaneous performance assessments. Published tests

TABLE 2
ROLE OF TEST TYPE AS A FUNCTION OF PURPOSE FOR ASSESSMENT, REPORTED BY GRADE AND SUBJECT
(in reliance percentages)

| Purpose | | Grade | | | | Subject | | | | Total Sample |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 2 | 5 | 8 | 11 | WR | SP | SC | MA | |
| | N | 57 | 58 | 58 | 55 | 58 | 61 | 50 | 59 | 228 |
| Diagnosing | OBJ* | 25% | 27 | 33 | 37 | 24 | 24 | 41 | 34 | 31 |
| | PUB | 20 | 25 | 12 | 13 | 19 | 14 | 15 | 21 | 17 |
| | ST PA | 24 | 23 | 27 | 35 | 37 | 26 | 23 | 22 | 27 |
| | SP PA | 32 | 26 | 28 | 15 | 20 | 35 | 21 | 24 | 25 |
| Grouping | OBJ | 25 | 27 | 32 | 32 | 20 | 24 | 36 | 38 | 29 |
| | PUB | 29 | 32 | 21 | 13 | 28 | 21 | 20 | 30 | 25 |
| | ST PA | 18 | 22 | 23 | 32 | 34 | 21 | 20 | 19 | 24 |
| | SP PA | 28 | 19 | 24 | 14 | 18 | 33 | 24 | 12 | 22 |
| Grading | OBJ | 29 | 36 | 43 | 48 | 34 | 33 | 46 | 44 | 39 |
| | PUB | 19 | 22 | 8 | 9 | 14 | 11 | 12 | 20 | 15 |
| | ST PA | 23 | 22 | 28 | 34 | 36 | 27 | 24 | 20 | 27 |
| | SP PA | 28 | 20 | 17 | 10 | 16 | 24 | 18 | 16 | 19 |
| Evaluating | OBJ | 30 | 35 | 36 | 39 | 31 | 33 | 44 | 35 | 35 |
| | PUB | 19 | 24 | 12 | 14 | 18 | 11 | 15 | 25 | 17 |
| | ST PA | 21 | 22 | 32 | 29 | 36 | 28 | 19 | 20 | 26 |
| | SP PA | 30 | 20 | 19 | 18 | 15 | 29 | 22 | 20 | 22 |
| Reporting | OBJ | 29 | 30 | 38 | 44 | 29 | 30 | 45 | 38 | 35 |
| | PUB | 22 | 29 | 14 | 10 | 20 | 14 | 17 | 26 | 19 |
| | ST PA | 25 | 23 | 30 | 31 | 35 | 28 | 22 | 23 | 27 |
| | SP PA | 26 | 18 | 18 | 14 | 17 | 28 | 18 | 13 | 19 |

* OBJ stands for teacher-made objective tests, PUB for published tests, ST PA for structured performance assessment and ST PA for spontaneous performance assessment.

0861e

are secondary. Reliance on objective tests increases with grade, as does reliance on structured performance assessment. Reliance on published tests fluctuates with grade, while the weight given to spontaneous performance assessment drops after grade 2. Across subjects, science and math teachers evaluate most heavily based on their own objective tests, while structured performance assessment is more important in writing and speaking.

And finally, when the purpose for assessment is reporting achievement results to parents, teachers rely most heavily on their own objective tests and structured performance assessment. Many of the same grade and school subject patterns referenced above appear here also. Objective and structured performance tests increase in importance as grade increases, while published and spontaneous performance assessment decrease in importance. Thus, math and science teachers weight their own objective tests most heavily, while writing and speaking teachers tend to use performance assessment.

## Concerns about Assessment

Type of Concern. Table 3 reports teachers' types of concern about different kinds of tests. The percentage of teachers selecting each category as her or his primary concern is reported.

Note that 28% of the total sample of teachers registered no concern about teacher-made objective tests. Thus, nearly three-quarters expressed some primary concern. By far the most common concern about teacher-made objective tests focused on test improvement, reflecting teachers' desire to improve their use of this kind of test. The other common concern is management, reflecting uneasiness with the amount of time required to manage this mode of assessment in the classroom. These

TABLE 3
SUMMARY OF TYPE OF CONCERN ABOUT ASSESSMENT
BY TEST TYPE, GRADE, AND SUBJECT
(in percent of respondents)

| Concern | Grade | | | | Subject | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 5 | 8 | 11 | WR | SP | SC | MA | Sample |
| N | 57 | 58 | 58 | 55 | 58 | 61 | 59 | 59 | 228 |
| **Teacher-made Objective Tests** | | | | | | | | | |
| No concern | 46% | 22 | 30 | 15 | 31 | 35 | 20 | 25 | 28 |
| Lack of information | 5 | -- | -- | -- | 3 | 2 | -- | -- | 1 |
| Competence | 2 | 2 | 2 | -- | -- | 2 | 2 | 2 | 1 |
| Time Management | 14 | 22 | 18 | 22 | 24 | 15 | 22 | 15 | 19 |
| Consequence | -- | 7 | 5 | 7 | 9 | 7 | 2 | 2 | 5 |
| Collaboration | 4 | 3 | -- | 7 | 3 | 2 | 2 | 7 | 4 |
| Improvement | 30 | 43 | 45 | 49 | 29 | 38 | 51 | 49 | 42 |
| **Published Tests** | | | | | | | | | |
| No concern | 41 | 31 | 42 | 38 | 38 | 37 | 32 | 45 | 38 |
| Lack of information | 5 | 9 | 5 | 9 | 10 | 8 | 10 | -- | 7 |
| Competence | -- | 3 | -- | -- | 2 | -- | -- | 2 | 1 |
| Time Management | 13 | 12 | 5 | 6 | 7 | 5 | 10 | 14 | 9 |
| Consequence | 20 | 17 | 14 | 33 | 19 | 15 | 28 | 22 | 21 |
| Collaboration | 2 | 3 | 4 | 4 | 3 | 2 | 2 | 5 | 3 |
| Improvement | 20 | 24 | 30 | 11 | 21 | 33 | 18 | 12 | 21 |
| **Structured Performance Assessment** | | | | | | | | | |
| No concern | 50 | 33 | 35 | 24 | 29 | 44 | 25 | 42 | 35 |
| Lack of information | 4 | -- | 4 | -- | 3 | 2 | 2 | -- | 2 |
| Competence | 4 | 4 | 7 | 2 | 3 | -- | 10 | 4 | 4 |
| Time Management | 20 | 21 | 16 | 26 | 21 | 12 | 27 | 25 | 21 |
| Consequence | 2 | 11 | 9 | 6 | 7 | 10 | 10 | -- | 7 |
| Collaboration | -- | 4 | 7 | 6 | 5 | 3 | 2 | 5 | 4 |
| Improvement | 20 | 28 | 23 | 38 | 31 | 29 | 25 | 25 | 27 |
| **Spontaneous Performance Assessment** | | | | | | | | | |
| No concern | 59 | 48 | 39 | 39 | 41 | 58 | 39 | 46 | 46 |
| Lack of information | 2 | 2 | 2 | 6 | 3 | 2 | 2 | 4 | 3 |
| Competence | 2 | 5 | 7 | 6 | 5 | 3 | 10 | 2 | 5 |
| Time Management | 9 | 10 | 2 | 4 | 10 | 3 | 6 | 5 | 6 |
| Consequence | 2 | 7 | 5 | 9 | 5 | 5 | 6 | 7 | 6 |
| Collaboration | 2 | -- | 2 | 7 | 2 | 2 | 4 | 4 | 3 |
| Improvement | 24 | 28 | 44 | 30 | 33 | 27 | 33 | 33 | 31 |

0861e

teachers do not tend to be concerned about a lack of information about these tests, their competence in using them, the student reactions to their use, or collaborating with others in using them. These patterns of concern vary with grade and slightly with subject. For example, about half of the second grade respondents expressed some concern, while 85% of the eleventh grade teachers did so. There is an increasing concern about quality and management of teacher-made objective tests as grade increases, and for math and science teachers in contrast to writing and speaking teachers.

Fewer teachers expressed specific concerns about published tests. About 40% of the total sample expressed no concern. Of those expressing some concern, most were uneasy about (1) student reactions and (2) test improvement. More eleventh grade teachers seem concerned about consequences than teachers at other grades. Beyond this, response patterns were generally stable across grades and subjects.

Expressions of concern about structured performance assessments were similar to those for teacher-made objective tests: improving quality and time management were most crucial. Some grade level trends appear, with indications that concern for improving such assessments and using them more effectively increases with grade level.

Spontaneous performance assessments elicit the fewest expressions of concern, with only half of the respondents reporting some concern. Most of these were concerned with improvement of the assessments. Again, the frequency of this concern seemed to gradually increase with grade level.

Reasons for Concern. After teachers indicated their primary concern, they were also asked to specify why that concern was primary for them.

The two most common types of concerns mentioned about teacher-made objective tests were improving test quality and time management. The reason for the teachers' concern about time required to develop and use their own tests is that it interferes with instructional time. Teachers who indicated uneasiness about the objective tests they developed and used posed such questions as: Are my tests effective? How can I make them better? Do they focus on students' real skills? Are they challenging enough? Do they aid in learning?

The two most frequent concerns about published tests related to students' reactions and improving the quality of test use. Those concerned about student reactions to published tests tended to view these tests as invalid, undependable, too long, etc. and thus anticipated that the tests were not helpful to students. Those concerned about improving test use see published tests as time-consuming, not matching their instruction, failing to reflect true student characteristics and generally not meeting important instructional needs such as identifying material to teach or reteach. For these reasons, they would like the tests revised and improved or would like to learn to use them more effectively. Published tests generated the most negative comments in respondents' expression of concerns. Many teachers see them as interfering with instruction.

Concerns about performance assessment--structured and spontaneous--dealt primarily with the desire to improve both the assessment and its use. Teachers' test quality concerns focused on accuracy of assessment, difficulty in defining levels of performance and the need to be objective. Test use issues reflected a desire to measure

growth, to challenge (but not intimidate) students, and to provide diagnostic information. Some were also concerned about the time demands of using performance assessments.

Classroom Performance Assessments

Seventy-eight percent of the teachers completing the questionnaire reported using structured performance assessments in their classrooms. Those 177 teachers responded to a series of questions which described their assessments. Results are presented by grade and subject in Table 4.

Responses to item 1 in Table 4 describe teachers' quality control procedures. Teachers were asked to indicate the percent of their performance asessments in which they include various procedures. On the average, teachers do the following in the majority of their assessments:

(Part A) specify a reason in their mind for assessment prior to testing

(Part C) inform students of their scoring criteria

(Part D) plan scoring procedures in advance

(Part E) define levels of performance assessment

On the other hand, less than half of the assessments include (B) written performance assessment criteria or (G) multiple performance observations before making a judgment. And finally, teachers seldom (F) rated performance without knowledge of the students' identity, or (H) cross-checked judgments about performance with other test scores.

There are some differences in responses across grades. For instance, as grade increases, so does the tendency to write down criteria and inform students of them, plan scoring procedures, define levels of

TABLE 4

DESCRIPTION OF PERFORMANCE ASSESSMENT BY GRADE AND SUBJECT

| | 2 | 5 | 8 | 11 | WRIT | SP | SCI | MATH | TOTAL |
|---|---|---|---|---|---|---|---|---|---|
| N | 38 | 51 | 41 | 47 | 46 | 46 | 38 | 47 | 177 |
| 1. In what percentage of all your STRUCTURED PERFORMANCE ASSESSMENTS do you | | | | | | | | | |
| A. Specify the reason for assessment in your own mind prior to conducting that assessment? | 79 | 82 | 88 | 85 | 87 | 86 | 76 | 84 | 83 |
| B. Write down scoring criteria before assessment? | 28 | 41 | 63 | 58 | 48 | 61 | 34 | 46 | 48 |
| C. Inform students of scoring criteria before assessment? | 35 | 62 | 73 | 78 | 66 | 73 | 57 | 55 | 63 |
| D. Plan actual scoring or rating procedures before assessment? | 57 | 61 | 76 | 75 | 70 | 73 | 57 | 68 | 67 |
| E. Clearly define levels of performance from adequate to inadequate before rating performance? | 48 | 58 | 76 | 69 | 66 | 69 | 56 | 60 | 63 |
| F. Conduct "blind" ratings of student products (i.e., rate performance without knowledge of who the respondent is)? | 8 | 14 | 28 | 23 | 10 | 12 | 32 | 23 | 18 |
| G. Observe and rate performance more than once before making a judgment? | 51 | 42 | 47 | 41 | 44 | 43 | 42 | 50 | 45 |
| H. Check your judgments against objective or published test scores before making a final decision? | 21 | 21 | 22 | 18 | 12 | 13 | 19 | 38 | 21 |
| 2. What percentage of all of your STRUCTURED PERFORMANCE TESTS involve the evaluation of | | | | | | | | | |
| Students doing things (Behavior)? | 52 | 49 | 55 | 55 | 32 | 67 | 56 | 57 | 53 |
| Products created by students? | 48 | 51 | 43 | 47 | 68 | 32 | 46 | 42 | 47 |

| E 4 (continued) | 2 | 5 | 8 | 11 | WRIT | SP | SCI | MATH | TOTAL |
|---|---|---|---|---|---|---|---|---|---|
| As you observe and rate performance, with what percentage of your assessments do you use the following procedures to record your judgments? | | | | | | | | | |
| A. Checklists (list of skills present or absent) | 31 | 30 | 35 | 35 | 35 | 40 | 24 | 21 | 33 |
| B. Rating Scales (continuum from good to poor quality performance) | 28 | 33 | 45 | 42 | 34 | 41 | 35 | 39 | 37 |
| C. Anecdotal Records (written descriptions of performance) | 23 | 23 | 35 | 33 | 26 | 41 | 20 | 25 | 28 |
| D. A Grade (in a record book) | 38 | 63 | 85 | 86 | 71 | 65 | 66 | 70 | 68 |
| E. Mental Notes (accumulated in memory over time) | 46 | 37 | 50 | 28 | 39 | 40 | 48 | 33 | 40 |
| What proportion of all of your STRUCTURED PERFORMANCE ASSESSMENTS do you score | | | | | | | | | |
| Holistically--scoring overall proficiency? | 26 | 28 | 29 | 22 | 27 | 15 | 32 | 31 | 26 |
| Analytically--scoring specific subskills? | 18 | 21 | 15 | 20 | 20 | 12 | 19 | 23 | 19 |
| Both holistically and analytically? | 54 | 51 | 55 | 60 | 50 | 75 | 49 | 44 | 55 |
| What proportion of your STRUCTURED PERFORMANCE ASSESSMENTS are conducted without students being aware that your are assessing them? | 40 | 25 | 13 | 13 | 17 | 22 | 26 | 23 | 22 |
| When rating students, do you always do the rating or do colleagues or the students themselves play a role? Indicate the appropriate percentage of ratings conducted by each potential rater listed below. | | | | | | | | | |
| A. I (the teacher) do the rating | 90 | 84 | 82 | 90 | 86 | 87 | 88 | 85 | 87 |
| B. Colleague rates student performance | 6 | 2 | 2 | 4 | 1 | 6 | 2 | 4 | 4 |
| C. Students rate each other's performance | 5 | 11 | 19 | 14 | 14 | 19 | 7 | 9 | 12 |
| D. Students rate their own performance | 13 | 19 | 16 | 10 | 15 | 18 | 12 | 13 | 15 |
| What proportion of your STRUCTURED PERFORMANCE ASSESSMENT results is interpreted primarily by comparing student performance to | | | | | | | | | |
| that of other students (norm referenced interpretation)? | 38 | 34 | 25 | 32 | 27 | 27 | 40 | 35 | 32 |
| specific preset standards of criteria of minimum acceptable performance (criterion referenced interpretation)? | 62 | 64 | 75 | 69 | 71 | 72 | 60 | 64 | 67 |

performance, and conduct blind ratings. Differences across subjects are less pronounced, but generally suggest that quality control activities do vary somewhat on this dimension also. For instance, teachers dealing with speaking assessment appear more likely to write down scoring criteria than others and are more likely to inform students of them than are math and science teachers. Further, it appears that science teachers are somewhat less likely to plan scoring procedures in advance of the assessment than are the others. Math and science teachers use blind scoring more frequently than their writing and speaking counterparts. And finally, teachers appear more likely to check their judgments against test scores when dealing with math in contrast to other subjects.

In the remaining items in Table 4, teachers further described characteristics of their structured performance assessments. Teachers reported that these assessments tended to be: equally divided between evaluations of process and product (item 2); recorded most frequently as a grade in the record book, and less frequently as mental notes, rating scales, checklists and anecdotal records (3); scored both holistically and analytically (4); conducted with the awareness of the student (5); based on teachers' judgments, with students rarely playing a role in self or peer assessment (6); and criterion referenced or based on pre-established standards of acceptable performance (7).

The data reported in Table 4 reveal some notable differences in test characteristics across grades and subjects. For instance, as grade increases, so does reliance on rating scales and grades. However, the use of unobtrusive assessment (5) decreases as grade increases. Comparing subjects, writing assessment is most frequently based on

product evaluation (presumably writing samples), while others are more process oriented. Speaking assessments use slightly more checklists and rating scales than others, while science assessors rely heavily on mental record keeping. Speaking assessments tend to be scored more completely (holistically and analytically) than others. All other characteristics are quite constant across subjects.

Relating Reliance Percentages Across Purposes. A correlational analysis was conducted to explore the question of whether teachers who rely heavily on one assessment procedure for one purpose tend to rely heavily on that same procedure for other purposes. Essentially, this is a follow-up analysis to the results presented in Table 2, where we saw that the weight given to any particular test type tended to vary only slightly across purposes. To verify this prior conclusion, we would need to find high correlations between weights assigned for the same test for different purposes. The results are presented in Table 5.

Since all 40 correlations reported are consistently quite large, a teacher's reliance on a particular assessment method appears to be somewhat stable across differing purposes. The average correlation between reliance indicators across purposes for teacher-made objective tests is .65 as it is for structured performance assessments. For the other two test types, the mean correlations were somewhat lower: .51 for published tests and .55 for spontaneous performance assessments. There are other notable patterns across matrices. For instance, in all four cases, the lowest correlation (average .44) is between the weight given to an assessment procedure for evaluating instruction and the weight given to that same procedure for instructional grouping. Also, in all

TABLE 5

CORRELATIONS AMONG IMPORTANCE RATINGS OF THE SAME TEST TYPE USED FOR DIFFERENT PURPOSES

| | Teacher-made Objective Tests | | | | Published Tests | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (1) | (2) | (3) | (4) |
| (1) Diagnosis | | | | | | | | |
| (2) Grouping | .61 | | | | .51 | | | |
| (3) Grading | .76 | .59 | | | .54 | .57 | | |
| (4) Evaluating | .60 | .44 | .71 | | .52 | .46 | .67 | |
| (5) Reporting | .74 | .57 | .81 | .69 | .54 | .63 | .67 | .70 |

| | Structured Performance Assessment | | | | Spontaneous P.A. | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (1) | (2) | (3) | (4) |
| (1) Diagnosis | | | | | | | | |
| (2) Grouping | .63 | | | | .62 | | | |
| (3) Grading | .82 | .56 | | | .62 | .45 | | |
| (4) Evaluating | .64 | .45 | .72 | | .48 | .42 | .48 | |
| (5) Reporting | .71 | .56 | .74 | .71 | .59 | .56 | .72 | .53 |

four cases, the highest correlations are found between the weight given when reporting achievement to parents and the weight given in grading (averaging .73), and between the weight given in grading and that given in diagnosing student strengths and weaknesses (.68).

## DISCUSSION AND CONCLUSIONS

From these results, we have selected five major issues for further analysis and discussion. These issues capture what we feel are the most important insights about classroom assessment to be derived from the data. In this section, we draw conclusions about (1) the use and importance of performance assessment in the classroom; (2) the stability of results across grades, subjects, and research contexts; (3) teachers' concerns about assessment, particularly with respect to improving test quality and use; (4) specific issues of assessment quality, including potential difficulties in classroom performance assessment procedures; and (5) actions needed to overcome some of the assessment problems.

### The Nature and Role of Performance Assessment in the Classroom

Our previous studies (Stiggins and Bridgeford, 1982) led to the conclusion that performance assessment--the observation and rating of student behavior and/or products--plays a key role in the day-to-day measurement of student achievement in the classroom. This study reinforces that conclusion. A large majority (177 of 228) of the teachers in this study report using structured performance assessment in the classroom. More important, the weights assigned to structured and spontaneous performance assessments show them to be heavily used modes of assessment in all five decision contexts explored. This appears to be true across the grades and school subjects examined. Our data indicate that performance assessment and teacher-made objective tests form the basis of most classroom assessment. Published tests play a secondary role. Teachers, moreover, have considerable confidence in their ability to make accurate observations and professional judgments; they express

comfort with performance assessment, and rely on it as a key method of judging students' learning. But the data also indicate that this confidence should not be confused with complacency. As we have seen, many teachers are sensitive to the fact that there may be problems in their assessments, are concerned about improving test quality and want to find ways to improve test use.

What are classroom performance assessments like? In one sense, they vary greatly across teachers and in another sense they are quite similar. The specific ingredients of the tests vary across subjects and to a certain extent across grades. Exercises, performance criteria, and student responses obviously vary as a function of school subject. However, the form of the assessment remains constant. Teachers evaluate both behaviors and products in approximately equal proportions. They tend to use prespecified standards (rather than student comparisons), to record assessment results with a grade in a record book, and not to involve students in performance ratings. Though most teachers know in advance why they are assessing--a key to quality assessment--some may fail to apply other quality control procedures to their performance assessments. We will explore this point in greater detail below.

Examining common characteristics of these assessments leads to the conclusion that performance assessment may not be used as effectively as is possible. For instance, students represent an untapped reservoir of performance raters, especially when teacher time is at a premium. Students can successfully rate their own and one another's performance and can learn a great deal from doing so (Spandel, 1981). For another example, recording systems other than grades often provide valuable and rich feedback to students. Checklists, rating scales, and anecdotal

records, for example, offer the detail often needed to describe performance and make careful assessments. The heavy reliance on grades seen in the data suggest that these alternatives are not being used to advantage.

Thus, results from this study confirm that performance assessment is an important assessment tool for teachers in the classroom. Results also indicate that the use of this assessment method could be enhanced and expanded.

## Stability and Change in Assessment Procedures

Within the pattern of relatively constant assessment methods, however, there are a few variations worthy of note. In this section, we explore the implications of those variations across, grade, subject and test type.

We found three interesting changes in assessment procedures as grade increases. First, the higher the grade level, the greater the tendency for teachers to report using their own assessments rather than published tests. Second, teachers' concern about assessment increases with grade level. And third, teacher's attention to quality control issues with performance assessments increases slightly with grade level. Levels of use of performance assessment as well as specific attributes of those assessments vary somewhat across grades. Thus, grade level appears to be an important variable in understanding classroom assessment. Elementary, junior high and high school environments differ in fundamental ways. The increased use of teacher-developed tests at higher grade levels might reflect the teacher's need to tailor tests to cover unique classroom objectives at higher levels. The reason for increased concern about

assessment across grade levels may relate to the increased importance placed on grades as a measure of student progress and success as grade increases. And increased attention to quality control may reflect the increased concern with accurately judging and grading students: clearly grades take on more importance as students advance in the school system, and can influence future decisions of students. These and other speculations deserve further consideration in future research.

Assessment procedures also differ as a function of school subject. This is to be expected and our data support this notion. Math and science teachers tend to rely more heavily on paper-and-pencil tests than do writing and speaking teachers. Speaking and writing teachers tend to use more performance assessments and the performance assessments they use tend to differ somewhat from those used by math and science teachers. Regardless, concerns about improving test quality and use tend to remain quite constant across subject.

We can also draw some conclusions, based on the data, about variations in assessment approach among teachers and for a given teacher. For instance, we have evidence that these teachers are relatively consistent in the assessment methods they use. They do not vary their testing methods very much as the purpose for assessment varies. This finding calls into question our conclusion in earlier studies that performance tests are instructional tools while objective tests are grading tools (Stiggins and Bridgeford, 1982). Both tests appear to play a role in both purposes. As these teachers described their levels of use, only a handful of the 228 teachers reported that they anticipated using or were preparing to use a new type of assessment

in the future. These teachers are not exploring new assessment approaches. This conclusion has implications for the action plans outlined below.

Teachers' Concerns About Assessment

At least three-quarters of the 228 teachers queried in this study expressed some concern about the assessments they used. Further, over half of the respondents indicated concern about each of the four assessment methods. Even when teachers reported relatively comfortable use of a given form of assessment, they were not reluctant to express a desire to improve their tests and the manner in which those tests are used. Their most frequently expressed concern involved improving the quality and use of assessments. Added to that, teachers frequently reported concern about their ability to effectively integrate assessment given the time constraints imposed by the classroom. Overall, teachers' responses in this study indicated concern about assessment quality and frustration at the lack of time available to deal more adequately with the problem.

But even more paradoxical and potentially troubling is the fact that although teachers are obviously concerned and many want to improve, at the same time (as cited above) these same teachers do not appear to be in the process of changing in ways that will improve their assessment methods. Clearly, many--though concerned--appear to lack the opportunity, time, means or motivation to revise their assessment approaches. We consider this dilemma further in addressing needed action programs.

The Extent of the Problem

Obviously, many teachers wonder about the effectiveness of the assessments they are using. But is there really reason to be concerned? Information on this issue from our data is limited but provides some insight. From the self-report data on quality control efforts in structured performance assessments, teachers' uneasiness may be justified. For example, in at least a third of the structured performance assessments conducted by these teachers, important assessment procedures appear not to be followed: students are not informed of performance criteria, scoring procedures are not planned in advance, and levels of performance (adequate to inadequate) are not defined before rating performance. Further, in over half of these assessments on the average, scoring criteria are not written down, judgments are based on a single observation, and performance ratings are not checked against other indicators, such as test scores. Finally, in an average of 40 percent of the structured performance assessments, teachers rely on mental record keeping. Since these practices can contribute significantly to the invalidity and/or unreliability of structured performance assessment results, there seems to be reason for concern.

Thus, the data suggest real problems. But caution is needed in interpreting these problems. The statistics presented above can be interpreted from a "glass half empty" or "glass half full" perspective. Pessimists say we have much to do. Optimists say much is already being done. Both are right. Many teachers do an excellent job of assessing, adhering to key aspects of quality control in the important assessments. In our discussions, interviews, and questionnaire responses, we found many very creative applications of performance assessment used in the

classroom, and there appears to be a strong foundation of good assessment present in many classrooms. We can build from that. Many teachers are not complacent. We can count on that. So, how do we proceed?

Moving Toward a Solution

Though the extent and depth of this problem is only suggested by these data, the problem is obviously significant. To deal with it, we propose a solution including four parts: (1) greater sensitivity to teachers' needs on the part of the measurement community; (2) more qualitative research on classroom assessment practices, (3) collaboration among teachers and (4) inservice training designed to meet teachers' needs. We have two key factors in our favor as we consider changes. First, our data on concerns suggest that many teachers are aware of the need to use assessment more effectively; they want to improve. Second, many teachers are strong assessors.

How can we use these factors to advantage? First and foremost, the measurement community must give greater attention to the classroom assessment needs of teachers. With a few notable exceptions, as a community of educators, we have only a limited understanding of the classroom assessment environment and teachers' most pressing assessment concerns. Evidence of this fact is presented in Table 6. We found that teachers rely on both observational assessment and teacher-made objective tests; published tests have considerably less influence on teachers. Yet, textbooks used in teacher training provide almost no instruction in the assessment methods most relevant for classroom use. Even more important, measurement research (as reported in professional journals) concentrates on assessment methods that have the least utility for

TABLE 6
RELATIVE IMPORTANCE OF TEST TYPE IN THE PROFESSIONAL LITERATURE AND IN TERMS OF TEACHERS' NEEDS

| Emphasis on: | In Texts[1] | In Research[2] | For teachers[3] |
|---|---|---|---|
| Teacher-made objective tests | 47% | 29% | 34% |
| Published tests | 47% | 62% | 19% |
| Performance Assessment | 6% | 9% | 47% |

1 Approximate percent of text pages on test construction and use in six introductory measurement test books: Ahman & Glock, 1971; Brown, 1970; Ebel, 1979; Gronlund, 1981; Mehrens and Lehmann, 1973; Noll, et al., 1979.

2 Approximate percent of articles dealing with those tests and test development in volumes 17, 18, 19 and 20 (1980, 81, 82, 83) of the Journal of Educational Measurement.

3 Reliance percentages summarized from Table 2, averaged across purposes and combined structure and spontaneous performance assessments.

0861e

teachers' decision making. As researchers, our focus must be redirected to include assessment methods and quality control issues in the classroom environment that affect student learning and instruction.

Second, we need more research on the classroom assessment needs of teachers. Extensive research on the role and use of standardized test scores in the classroom has certainly played an important role in helping us deal with some key assessment problems. But the time has now come to move to a new emphasis; namely, understanding the role of strategies such as teacher observation in classroom assessment. The research reported here represents a small but potentially useful step in that direction. We might also follow the lead of Good and Brophy (1978), who have provided teachers with systematic strategies for observing in the classroom.

Third, teachers who are competent assessors are another vital training resource which must be tapped. Results of this study suggest that teachers who rely most heavily on performance assessments tend to use such tests somewhat more carefully than those who use them less. Teachers with assessment skill can assist their colleagues. Previous research revealed that teachers regard colleagues as one of the two most important sources of assessment ideas (Stiggins and Bridgeford, 1982). Yet this study revealed little or no collaboration among teachers in test use. These two findings identify a valuable source of ideas that is not being tapped. Why? Because there is no time, encouragement or planning to do so. Test quality may be readily improved by encouraging and promoting collaboration in assessment.

Greater awareness of the classroom assessment environment and its demands can form the basis for another important element in our plan of action: relevant training for teachers. Based on the textbooks examined in Table 6, current and past training is out of balance. Further, a large proportion of teachers have had no measurement training at all (Stiggins and Bridgeford, 1982; Coffman, 1983). Many teacher preparation programs (graduate and undergraduate) do not require measurement training and many teachers avoid it, given a choice. One reason for this avoidance is that our training fails by reputation to meet important teacher needs.

As we design and develop training that is more relevant to teachers' classroom assessment needs, all available resources must be tapped. For instance, graduate and undergraduate teacher preparation courses continue to offer an opportunity for relevant training. Perhaps the student teaching experience could be structured to deal directly with classroom assessment issues. But inservice training, structured to meet teachers' assessment needs, provides the greatest opportunity for impact. The key to success in both settings will not be to present more "strategies to interpret standardized test scores." Kellaghan and others (1982) have shown these have little impact on teachers' testing practices. Instead, training must focus on real teacher needs and provide guidance in quality control for all teacher-made tests, including those based on observations and subjective judgments.

## REFERENCES


Ahmann, J.S. and M.D. Glock, Evaluating Pupil Growth (fourth edition). Allyn-Bacon, 1971.

Airasian, P.W., et al. Proportion and direction of teacher rating changes of pupil progress attributable to standardized test information. Journal of Educational Psychology, 1977, 69(6), 702-709.

Brown, F.G. Principles of Educational and Psychological Testing, Dryden, 1970.

Burstein, L. A word about this issue. Journal of Educational Measurement, 1983, 20(2), pp. 99-101.

Calfee, R.C. and P.A. Drum. How the Researcher Can Help the Reading Teacher with Classroom Assessment. Unpublished manuscript, Stanford University, Stanford, CA, 1976.

Coffman, W.E. Testing in the Schools: A Historical Perspective. A paper presented at the Center for the Study of Evaluation Annual Invitational Conference, UCLA, 1983.

Ebel, Robert L. Essentials of Educational Measurement (third edition). Prentice-Hall, 1979.

Fitzpatrick, R. and E.J. Morrison. Performance and product evaluation. In R.L. Thorndike (Ed.) Educational Measurement (second edition). Washington, D.C.: American Council on Education, 1971, 237-270.

Glaser, R. and D.J. Klaus. Proficiency measurement: Assessing human performance. In R.M. Gagne (Ed.) Psychological Principles in System Development. New York, NY: Holt, Rinehart and Winston, 1962. 419-474.

Good, T.L. and J.E. Brophy. Looking in Classrooms (second edition). New York, NY: Harper and Row, 1978.

Goslin, D.A. Teachers and Testing. New York, NY: Russell Sage Foundation, 1967.

Gronlund, M.E. Measurement and Evaluation in Teaching (fourth edition). McMillan, 1981.

Gullicksen, A. Survey Data Collected in Survey of South Dakota Teacher's Attitudes and Opinions toward Testing. University of South Dakota, 1982.

Hall, G.E., A.A. George and W.L. Rutherford. Measuring Stages of Concern about the Innovation: A Manual for Use of the SoC Questionnaire. Austin, TX: Research and Development Center for Teacher Education, University of Texas, 1977.

Herman, J. and D.W. Dorr-Bremme. Assessing Students: Teacher's Routine Practices and Reasoning. Paper presented at the annual meeting of the American Educational Research Association, New York, 1982.

Kellaghan, T., G.F. Madaus and P.W. Airasian. The Effects of Standardized Testing. Boston, MA: Kluwer-Nijhoff, 1982.

Lindquist, E.F. Preliminary considerations in objective test construction. In E.F. Lindquist (Ed.) Educational Measurement, Washington, D.C.: American Council on Education, 1951.

Lortie, D. School Teacher. Chicago, IL: University of Chicago Press, 1975.

Mehrens, W.A. and I.J. Lehmann. Measurement and Evaluation in Education and Psychology. New York, NY: Holt, Rinehart and Winston, 1973.

Noll, V.H., D.P. Scannell, and R.C. Craig. Introduction to Educational Measurement (fourth edition). Houghton Mifflin, 1979.

Rudman, H.E., et al. Integrating Assessment with Instruction: A Review (1922-1980). Research Series #75. East Lansing, MI: College of Education, Michigan State University, 1980.

Ryans, D.G. and N. Fredericksen. Performance tests of educational achievement. In E.F. Lindquist (Ed.) Educational Measurement, Washington, D.C.: American Council on Education, 1961, 455-493.

Salmon-Cox, L. Teachers and standardized achievement tests: What's really happening? Phi Delta Kappan, May 1982, 631-634.

Scates, D.E. Differences between measurement criteria of pure scientists and of classroom teachers. Journal of Educational Research, 1943, 37, 1-13.

Spandel, V. Classroom Applications of Writing Assessment, Portland, OR: Northwest Regional Educational Laboratory, 1981.

Sproull, L. and D. Zubrow. Standardized testing from the administrative perspective. Phi Delta Kappan, May 1982, 628-631.

Stetz, F. and M. Beck. Comments from the Classroom: Teachers' and Students' Opinions of Achievement Tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Stiggins, R.J. Evaluating Students Through Classroom Observation: Watching Students Grow. Washington D.C.: National Education Association, in press.

Stiggins, R.J. and N.J. Bridgeford. Final Research Report on the Nature, Role and Quality of Classroom Performance Assessment. Portland, OR: Northwest Regional Educational Laboratory, 1982.

Yeh, J. Test Use in the Schools. Los Angeles, CA: Center for the Study of Evaluation, UCLA, 1978.