Complete Work- Teacher Training in Measurement and Assessment Skills

Jane Close Conoley
*University of California - Riverside*

Steven L. Wise
*University of Nebraska-Lincoln, wisesl@jmu.edu*

Follow this and additional works at: [https://digitalcommons.unl.edu/burosteachertraining](https://digitalcommons.unl.edu/burosteachertraining)


This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Teacher Training in Measurement and Assessment Skills by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Teacher Training in Measurement and Assessment Skills
Teacher Training in Measurement and Assessment Skills

Edited by

STEVEN L. WISE
University of Nebraska-Lincoln

BUROS INSTITUTE OF MENTAL MEASUREMENTS
University of Nebraska-Lincoln
Contents

Preface vii


2. Teacher Training in Assessment: Overcoming the Neglect Richard J. Stiggins 27

3. The Development of Standards for Teacher Competence in Educational Assessment of Students James R. Sanders and Suzanne R. Vogel 41

4. Some Thoughts on Grading Systems and Practices James S. Terwilliger 63

5. Teachers' Assessment of Students: Roles, Responsibilities, and Purpose Donna Campbell 97


7. Measurement Training in Nebraska Teacher Education Programs Steven L. Wise and Leslie E. Lukin 187
8. Thoughts on the Relationship Between Measurement Knowledge and Teacher Effectiveness
   Jack J. Kramer

   Author Index

   Subject Index
The Buros-Nebraska Symposia on Measurement and Testing were developed to provide a forum for discussion of issues relevant to effective use of tests and measurement. The topic of the 1989 symposium was “Are Our School Teachers Adequately Trained in Measurement and Assessment Skills?” This topic was selected because of a growing interest in the knowledge and skill levels of teachers in measurement and assessment. There has been considerable debate regarding the amount and types of skills needed for effective evaluation of student achievement.

This volume, like previous ones in this series, reflects many of the papers presented at the symposium as well as additional invited chapters that complement the chapters based on the symposium presentations. An attempt was made to broadly address, from a variety of perspectives, the measurement issues encountered by teachers.

The first chapter in the volume is authored by Arlen Gullickson and is titled “Matching Measurement Instruction to Classroom-Based Evaluation: Perceived Discrepancies, Needs, and Challenges.” Dr. Gullickson discusses previous research on the measurement training of teachers and provides a number of recommendations for improving the ways in which teachers are trained.

Richard Stiggins is the author of the second chapter, “Teacher Training in Assessment: Overcoming the Neglect.” He asserts that
there is a mismatch between what teachers need to know about assessment and what they are taught in teacher education programs. Dr. Stiggins makes recommendations for eliminating this mismatch.

“The Development of Standards for Teacher Competence in Educational Assessment of Students” is the title of the next chapter authored by James Sanders and Suzanne Vogel. They discuss the chronology of the Standards, beginning with the growing concern regarding the quality of student assessments and their use by educators and resulting in the Standards, which are included in the chapter.

James Terwilliger’s chapter, “Some thoughts on Grading Systems and Practices,” addresses the role of student grades in the assessment process. Dr. Terwilliger examines this issue from philosophical, theoretical, and empirical standpoints.

An alternative perspective on assessment issues in schools is provided in Donna Campbell’s chapter, “Teachers’ Assessment of Students: Roles, Responsibilities, and Purpose.” She discusses the philosophical underpinnings of current assessment beliefs and practices, and suggests a thoughtful reconceptualization of the role and meaning of assessment.

In the next chapter, “Teacher’s Testing Knowledge, Skills, and Practices,” Ronald Marso and Fred Pigge provide a review of the measurement literature on teachers’ testing knowledge and skills. Drs. Marso and Pigge also make a number of suggestions regarding how the measurement profession can improve current training strategies.

In “Measurement Training in Nebraska Teacher Education Programs,” Steven Wise and Leslie Lukin present the results of a study investigating the measurement training provided by one state’s teacher education programs. In addition, they report the findings of a survey of state teachers regarding their beliefs and attitudes about the adequacy of their measurement training.

In the final chapter, “Thoughts on the Relationship Between Measurement Knowledge and Teacher Effectiveness,” Jack Kramer reviews the other papers in this volume and provides an additional perspective. Dr. Kramer asserts that a greater understanding is needed of how teachers actually measure behavior in the classroom, and how specific measurement practices influence student achievement.

Teaching requires a complex set of measurement and assessment skills. These skills include the administration and interpretation of standardized tests, the ability to make rapid in-classroom assessment
of student understanding and progress, the measurement of student achievement, assignment of grades, and the ability to explain assessment results to parents. Moreover, the diversity of skills needed appears to be increasing. It is hoped that this volume will contribute to a greater understanding of the needed skills and how best to prepare teachers for a life of assessment in the classroom.

Steven L. Wise
Matching Measurement Instruction to Classroom-Based Evaluation: Perceived Discrepancies, Needs, and Challenges

Arlen R. Gullickson
University of South Dakota

Teacher knowledge about measurement, testing practices, and what teachers should be taught have been recurrent topics of concern in the past two and a half decades. Conant (1963) first captured measurement professionals' interest with his book *The Education of American Teachers*. That book stimulated a National Council on Measurement in Education (NCME) symposium regarding the implications of his recommendations for measurement instruction. Papers presented at the meeting were published in the first volume of the *Journal of Educational Measurement (JEM)*. Thus, in a sense, the issue before us is one of the most enduring in the NCME organization.

Since that initial volume of the *JEM*, the issues confronting us today have surfaced repeatedly. Work by the authors Goslin (1967), Mayo (1964, 1967), and Rudman et al. (1980) stands as perhaps the most significant early efforts. Goslin and Mayo tended (a) to highlight the importance of teaching teachers about testing, (b) to define the content emphasized in measurement courses, and (c) to identify the major differences in teachers' and measurement professionals' perceptions regarding what should be emphasized in measurement
courses. The review of literature by Rudman et al. in general served to heighten concerns about the measurement practices that take place in the classroom. Their review cites numerous individuals who have argued that teachers are not sufficiently knowledgeable, that the wrong content is being emphasized in teaching teachers, and that measurement specialists are not sufficiently knowledgeable about teacher testing practices. They put the issue in perspective with the following statement:

A troublesome aspect in this area is the paucity of descriptive material compared to the abundance of prescriptive articles, essays and the like dealing with the specifics of how teachers used test results in their classroom. When coupled with the information supplied by Beck and Stetz (in press) concerning the relatively inaccurate perceptions of measurement specialists who write about teacher testing behavior, positive conclusions about how teachers use tests can be only fragile speculations at best. (p. 20)

Since 1980, numerous studies have been conducted. Teachers have been surveyed and interviewed to learn about teacher attitudes and evaluation practices, teachers and students have been observed in the classroom, teacher certification requirements for educational measurement (or lack thereof) have been identified and noted, and professors of educational measurement courses along with elementary and secondary teachers have been surveyed to assess what is and should be taught in these measurement courses.

These more recent studies present a deepening concern about the knowledge of teachers, the evaluation practices that teachers employ in the classroom, and the measurement content and concepts emphasized in the preparation of teachers. Together the studies have stimulated substantial interest in the measurement preparation teachers should receive. Most notably, the NCME has initiated a task force of teachers, administrators, and measurement specialists to generate standards for teacher preparation in educational assessment of students.

If the measurement profession is to set standards for the measurement and evaluation preparation of teachers, there must first be agreement regarding the content to be taught. This paper focuses on that issue of content. Specifically, the issue broached is: What content should be provided in teachers’ undergraduate preparation in order to serve them best as they begin to teach?

The stage for this discussion is set by recounting in some detail findings from four separate but related survey efforts, which individually addressed (a) teacher attitudes toward testing, (b) teacher
testing and evaluation practices in the classroom, (c) teacher beliefs about what measurement topics and concepts should be taught at the preservice level, and (d) professors' perceptions of the actual characteristics of undergraduate measurement courses. (Articles by Gullickson, 1982, 1984a, 1984b, 1985, 1986; Gullickson & Ellwein, 1985; and Gullickson & Hopkins, 1987, provide details regarding the samples, instruments, and methods employed in those surveys.) The composite findings are intended to clarify:

- the way in which teachers view and use measurement and evaluation in their classrooms
- the context within which measurement is taught at the undergraduate college level—and content presently emphasized in those courses
- perceived strong differences of opinion between teachers and professors regarding what should be taught in the undergraduate measurement and evaluation courses

These three factors (facets) will then serve as a backdrop for addressing the central issue of what should be taught in the preservice measurement and evaluation course.

TEACHER ATTITUDES AND PRACTICES

The first two survey efforts, those directed to elementary and secondary teachers, sought primarily to learn about teacher testing practices. That focus was in concert not only with the author's measurement orientation toward measurement instruction but was also in tune with most professionals who talked about the preparation of teachers. It seems that routinely the course is referred to as "Tests and Measurement.”

The surveys were conducted in the early 1980s. At that time the popular press raised questions that suggested teachers were opposed to tests. Instead of being opposed to tests, the surveys revealed that teachers view tests, particularly teacher-made tests, as important instructional tools. Teachers reported that tests provide direct instructional benefit to them by helping to focus teaching, by providing feedback on instruction, and by providing feedback on student progress.

Teachers also reported they view tests to be of direct benefit to students. That is, they perceive the act of taking a test to be a learning experience for students. But, more broadly, they believe that tests motivate students to study, create competition among students, improve student interaction, have an important effect on student
self-concept, and do not negatively affect student attitudes toward the course.

Besides those direct statements of importance, other indirect factors lead to the same conclusions. For example, a large majority of teachers use tests, give tests frequently, and spend a great deal of time engaged in the testing process. As can be expected, not all teachers view and use tests in the same way. Thus, there are patterns in each of these factors that can help us to better understand the special relationship between teachers and their tests.

Eighty-nine percent of elementary teachers report using tests, whereas virtually all secondary teachers (99%) report such use. In using tests, they argue that it is better to give more frequent short tests than it is to give long tests infrequently. Thus, it is not surprising that 16% claim to test daily, 95% report weekly use of tests, and 98% report at least biweekly use.

The typical teacher devotes a considerable portion of personal preparation time and class time to the testing program. If one conservatively estimates that one test per course is given every other week, the information provided in Table 1 suggests that for each

<table>
<thead>
<tr>
<th>Test Activity</th>
<th>Elem</th>
<th>Jr</th>
<th>Sr</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Development</td>
<td>30</td>
<td>60</td>
<td>60</td>
<td>30</td>
</tr>
<tr>
<td>Pre-Review</td>
<td>30</td>
<td>40</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>Test Administration</td>
<td>30</td>
<td>35</td>
<td>45</td>
<td>35</td>
</tr>
<tr>
<td>Test Correcting</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>Post-Review</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Total Time</td>
<td>125</td>
<td>190</td>
<td>230</td>
<td>190</td>
</tr>
</tbody>
</table>

*Note.* Column Header Abbreviations are: Elem = Elementary, Jr = Junior High, Sr = Senior High
course, at least one class period per week is devoted to the activities of pretest review, test taking, and posttest feedback. Additionally, the teacher devotes up to another hour per week preparing and correcting each administered test. Those figures argue strongly that the typical teacher spends at least 20% of his or her time on testing activities; more likely this teacher spends over a third of his or her time in such activities.

Teachers view themselves as being in charge of the testing done for instructional purposes. They decide what tests to give, when to give the tests, and what to evaluate. The actual role of tests in the classroom tends to vary by test type, by grade level, and even by curriculum. Although the teacher-made objective test is the dominant testing practice across all grades and curricula, essay tests play a relatively prominent role at the senior high level, as do standardized objective tests and quizzes at the elementary level.

Teachers indicate a preference for creating their own test items, but as Table 2 shows, they do use other sources as well, principally textbook publisher-prepared items (see Green & Stager, 1986 for supporting data). Consistent with teachers’ preference for objective tests, Table 3 shows objective items, particularly short answer/completion, as the most common item type.

**Table 2.** Teacher Reported Primary Sources of Test Items for Tests They Use

<table>
<thead>
<tr>
<th>Item Source</th>
<th>Elem</th>
<th>Jr</th>
<th>Sr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=92</td>
<td>n=88</td>
<td>n=129</td>
</tr>
<tr>
<td>Self</td>
<td>86</td>
<td>97</td>
<td>96</td>
</tr>
<tr>
<td>Publisher of Text</td>
<td>75</td>
<td>61</td>
<td>47</td>
</tr>
<tr>
<td>Other Teachers</td>
<td>9</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>Other Published Items</td>
<td>21</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>Other</td>
<td>7</td>
<td>11</td>
<td>9</td>
</tr>
</tbody>
</table>

*Note.* Teachers were asked to select all options which serve as primary item sources. All values are reported as percents. Column Header Abbreviations are: Elem = Elementary, Jr = Junior High, Sr = Senior High
Table 3. Percent of Teachers Who Normally Use the Respective Item Types in their Tests

<table>
<thead>
<tr>
<th>Item Type</th>
<th>El</th>
<th>Jr</th>
<th>Sr</th>
<th>Sci</th>
<th>SS</th>
<th>LA</th>
<th>Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Answer/Completion</td>
<td>94</td>
<td>87</td>
<td>128</td>
<td>109</td>
<td>104</td>
<td>96</td>
<td>309</td>
</tr>
<tr>
<td>Matching</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>92</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>75</td>
<td>86</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>77</td>
</tr>
<tr>
<td>True-False</td>
<td></td>
<td></td>
<td></td>
<td>63</td>
<td>79</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>Essay</td>
<td>31</td>
<td>66</td>
<td>73</td>
<td>48</td>
<td>65</td>
<td>61</td>
<td></td>
</tr>
</tbody>
</table>

Note. Percentages are provided for the total group if there were no significant differences (p < .05) across grade and curriculum, or by grade and/or curriculum when significant differences existed for the respective groups. The column header abbreviations are: El = Elementary, Jr = Junior High, Sr = Senior High, Sci = Science, SS = Social Science, LA = Language Arts, and Tot = Total. The sample size for each group is provided directly below the column header.

Teachers’ choice of objective items bodes well for providing comprehensive content coverage, but not necessarily for test quality. Measurement professionals (e.g., Gronlund, 1985) argue that item types such as short answer and matching do not effectively measure higher cognitive levels. Teachers themselves endorse essay tests rather than objective tests as a means to measure higher cognitive levels. They believe essay tests (a) better evaluate higher cognitive level learning objectives than do objective tests and (b) in general provide a better evaluation of student learning than can be achieved through objective items. Thus, both groups appear to have some reservations about teachers’ preferences in item types.

The fact that a high proportion of teachers regularly uses item types designed to assess lower cognitive skills does not necessarily mean that their tests do not adequately measure higher order thinking skills. It does, however, suggest such a possibility. Indeed, other research (Carter, 1984; Fleming & Chambers, 1983; Stiggins, Griswold, & Wikelund, 1989) directly substantiates that teachers’ tests tend to focus on lower order thinking skills (recall of facts, etc.).

Just as teachers write their own tests, so do they administer, score, and grade them. Several aspects of teacher practices in these regards bear description. First, teachers correct and return tests quickly, almost always within 2 days. Second, teachers state that they use a criterion reference basis for grading tests. Third, teachers do little in the way of formal test analysis. Fourth, regardless of whether
individual tests are graded or just the total score on the test is returned to the student, test results play a prominent but not exclusive role in grading the student (Stake & Easley, 1978; Haertel, 1986).

Although the quick return can be considered a plus, the actual scoring and analysis process cannot. The information that teachers provide about their scoring and analysis practices suggests that the analysis is severely limited. That is, for the typical teacher, formal analysis includes only simple scoring, grading, and frequency counts of test scores.

Teachers' failure to more thoroughly analyze student test results may be due to lack of skill or lack of time. Teachers claim they know how to "item-analyze" their tests, but indicate such analysis is not practical. However, the inconsistencies in teacher responses to options on test analysis questions suggest teachers do lack the requisite skills.

Regardless, test correction and scoring constitute the only activities the typical teacher takes to assess instructional quality, to assess test quality, and to prepare feedback for the students. As a result, the standard fare for review of test results can be little more than a token statement about the distribution of test scores and a review of items selected by either the students or the teacher.

Given those limiting factors, the reviews cannot provide a clear perspective of which objectives were obtained by either individual students or the class in general. Thus, the review cannot adequately serve either the formative purposes for student instruction or formative purposes for revision of instruction.

Just as teachers' failure to fully analyze test results limits the instructional opportunities, so does it limit test improvement options. If tests or test items are reused, then an item analysis is helpful in detecting and correcting item flaws. Most teachers (84%) do reuse their tests, either in total or part. That reuse without attention to item analysis suggests teachers' tests do not significantly improve in quality over time.

An additional disquieting aspect of test quality comes in the form of a discrepancy between what teachers state that tests should be and what teachers state tests actually measure. Teachers state that tests should (a) be competency based, not norm based; and (b) measure learning in the target area, not just material explicitly assigned or covered in class. However, (c) they also believe the content of the test should emphasize the same material emphasized in class (their instructional emphases). These indicators suggest that the teacher-prepared test should fit the teacher's specified curricular objectives. Despite these expectations, teachers also report that they anticipate
their tests assess just 75% of that territory. Additionally, although the teachers argue that test results are a good indicator of how well the student has learned the material, they are not willing to stipulate that the test results are a good indication of how well the student will be able to apply what has been learned.

This teacher insecurity about the quality of tests is evident in the grading process. Teachers do view tests as an administrative necessity in justifying student grades. But, although virtually all teachers obtain a total score for each test and the strong majority (75%) do grade all or most of their tests, teachers argue that tests should not be used as the sole determiner of student grades. In fact, the typical teacher surveyed was not even willing to argue that tests should be used as the primary basis for assigning student grades. (That finding is contradicted by Haertel, 1986, p. 18. He found teachers in general did indicate the “unit test or midterm performance” to be the most important single factor in determining the student’s course grade.)

If teachers do not view tests as the primary basis for grades, what do they use in addition to tests? That question was not addressed directly, but the role teachers give to other evaluation techniques does give some insight into probable other sources. In particular, teachers report that student work products, teachers’ perception of student understanding through class discussion, and even student deportment all play a significant role in the overall evaluation process and the grading process in particular (see also Haertel, 1986).

Altogether, teachers appear to value tests as instructional tools and use them frequently. However, despite the teachers’ reported comfort with their testing skills, the survey results suggest numerous deficiencies both in their tests and in their testing skills. In particular, the tests appear to focus on lower cognitive skills and do not assess a substantial proportion of the teachers’ objectives. Further, the test analysis and feedback patterns suggest that teachers’ tests do not serve formative evaluation purposes.

MEASUREMENT INSTRUCTION AT THE UNDERGRADUATE LEVEL

Results from the survey of elementary and secondary teachers suggest strongly that teachers do not gain their knowledge of testing and evaluation practice from college courses. Naively, the author of the survey assumed that all teachers take measurement courses as a part of the preservice measurement preparation. Thus, the results of the survey initially were interpreted as an indictment of measurement courses.
Although it may be true that teachers do not view college measurement courses positively, an alternative explanation for teacher responses is that many teachers have had only a minimal exposure to educational measurement in their preservice courses. In fact, Noll (1955, p. 88) reported, "In sum, it may be said that a course in measurement for any teacher’s, administrator’s, or counselor’s certificate is a comparatively rare requirement, and even recommendation of such a course as an elective is not common." That condition has improved, but still, the measurement preparation of teachers is variable and tends to be minimal (Schafer & Lissitz, 1987; Haertel, 1986).

The survey of professors revealed that both course content and method of instruction vary substantially from college to college. A strong majority of colleges (71%) report that they offer an undergraduate course in educational measurement. Of these colleges, three fourths indicate the course is required. Thus, in roughly half the colleges, all preservice teachers must take an educational measurement course. In those schools where the course is optional, it is taken by a small portion of the students, typically 25% or fewer. The remaining students, those not taking a course, typically received some measurement instruction in the context of other courses (e.g., educational methods or educational psychology).

Students take the measurement course prior to student teaching, and in that course they receive a blend of theoretical and practical information. Professors indicate that they give both theory and practice a strong role in their instruction, with lecture/discussion taking about 50% of class time and student activities taking another 40% of the class period.

Eighty-two percent of the professors teaching the course have a doctorate, and all reported having at least a master’s degree. Most (74%) professors report their degree preparation, either as a major or minor, to be in an educational measurement-related area.

The professors report being experienced in education. Ninety-three percent report having taught at the elementary or secondary level, and they report substantial collegiate-level teaching experience as well.

Despite such experience, many of the educational measurement and evaluation professors are not formally a part of the curriculum and instruction discipline. Rather, they tend to come from other departments, such as educational psychology or statistics. In fact, for this course the use of adjunct professors or professors from outside education (e.g., psychology) appears to be fairly common.
There is little indication that the measurement and evaluation course is tied integrally to individual discipline areas. Instead, what appears to be the more common pattern is that the course simultaneously serves students from all discipline areas. Given the broad spectrum of students served, and the difficulty of finding examples that adequately serve all discipline areas, the course can be expected to focus on general principles of measurement without special emphasis being given to the techniques used most frequently either in the respective disciplines or at different grade levels.

**The content of undergraduate measurement courses.** To address the issue of course content, professors were presented with a list of 67 topics divided into the following eight categories:

1. General assessment information, which included items related to:
   a. Sources of aid in interpreting and using assessment information
   b. Selection and use of standardized and publisher-prepared tests
2. Preparing examinations, including:
   a. General development concerns
   b. Item selection and construction
3. Administering and scoring tests
4. Employing other evaluative devices
5. Computing and interpreting statistical data
6. Using test results for planning (formative evaluation) purposes
7. Using test results for summative evaluation purposes
8. Testing and the law—legal challenges to test practices

Professors were asked to rate the actual emphasis they personally gave to each of the topics. When the results were viewed by category, two topics—statistical analyses and exam preparation—received substantially higher ratings than did the other categories. Similarly, two topics, employing other evaluative devices and legal issues, were rated as receiving much less attention than the other areas (research by Stiggins & Conklin, 1988, provides substantiating evidence regarding non-test evaluation techniques). See Table 4 for a breakdown of emphasis by category.

Those findings suggest a clear, strong emphasis on testing with greatest emphasis given to creating, analyzing, and interpreting tests. In particular, it is noteworthy that professors designate non-test activities as being given very little emphasis.
1. MATCHING MEASUREMENT INSTRUCTION

Table 4. Means and Standard Deviations for Teachers and Professors

<table>
<thead>
<tr>
<th>Scale</th>
<th>Teachers (n=360)</th>
<th>Professors (n=24)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>s</td>
</tr>
<tr>
<td>1. General</td>
<td>3.24</td>
<td>0.83</td>
</tr>
<tr>
<td>2. Prep. of Exams</td>
<td>3.47</td>
<td>0.74</td>
</tr>
<tr>
<td>3. Admin./Scoring</td>
<td>3.39</td>
<td>0.91</td>
</tr>
<tr>
<td>4. Nontest</td>
<td>3.42</td>
<td>0.79</td>
</tr>
<tr>
<td>5. Statistics</td>
<td>2.78</td>
<td>0.93</td>
</tr>
<tr>
<td>6. Formative Eval.</td>
<td>3.58</td>
<td>0.80</td>
</tr>
<tr>
<td>7. Summative Eval.</td>
<td>3.48</td>
<td>0.80</td>
</tr>
<tr>
<td>8. Law</td>
<td>2.69</td>
<td>1.24</td>
</tr>
</tbody>
</table>


Measurement instruction emphases: A contrast of teacher and professor perspectives. Elementary and secondary teachers were presented with the same list of content emphases that professors rated. However, where professors were asked to rate emphases given to the topics, teachers were asked to rate the emphasis they believed should be given to the respective topics.

When compared with professor ratings, results of this survey show one area of strong agreement and at least two areas of strong disagreement. Professors and teachers strongly agree that test development issues are a high priority. But, although professors give greatest emphasis to statistical analyses, teachers desire little emphasis on that category. Just the opposite is true regarding the category of other evaluative devices. There teachers desire a strong emphasis, but professors give it little emphasis.

Table 5 provides a different and, in some respects, a more detailed perspective of similarities and differences in teacher and professor priorities. That table presents the top 20 priorities for both teachers and professors. The left column of this table was created by selecting and grouping the 20 topics teachers value most highly. Similarly, the right column represents the 20 topics professors emphasize most. The
Table 5. The 20 Content Priorities Which Teachers and Professors Respectively Rate Most Highly for Undergraduate Level Educational Measurement Courses

<table>
<thead>
<tr>
<th>Teacher Desired Emphases</th>
<th>Emphasis Given by Professors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Preparation</strong></td>
<td></td>
</tr>
<tr>
<td>Preparation of exams</td>
<td>Preparation of exams</td>
</tr>
<tr>
<td>Defining course objectives</td>
<td>Defining course objectives</td>
</tr>
<tr>
<td>Determining appropriateness of test content for specific classes</td>
<td>Defining skill and taxonomy Levels</td>
</tr>
<tr>
<td>Item selection and construction</td>
<td>Item selection and construction</td>
</tr>
<tr>
<td>Writing test items</td>
<td>Writing test items</td>
</tr>
<tr>
<td>Writing objective items</td>
<td>Writing objective items</td>
</tr>
<tr>
<td>Writing subjective test items</td>
<td>Writing subjective test items</td>
</tr>
<tr>
<td><strong>Test Statistics and Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Administering and scoring tests</td>
<td>Standard scores and the normal distribution</td>
</tr>
<tr>
<td>Scoring Tests</td>
<td>Measures of central tendency and variability</td>
</tr>
<tr>
<td></td>
<td>Computing and interpreting statistical data</td>
</tr>
<tr>
<td></td>
<td>Correlations and reliability coefficients</td>
</tr>
<tr>
<td></td>
<td>Percentages and percentiles</td>
</tr>
<tr>
<td></td>
<td>Transforming raw scores</td>
</tr>
<tr>
<td><strong>Formative and Summative Use of Tests</strong></td>
<td></td>
</tr>
<tr>
<td>Interpreting test profiles to identify pupil strengths and weaknesses</td>
<td>Using test results for planning (formative evaluation) purposes</td>
</tr>
<tr>
<td>Identifying gifted pupils or slow learners</td>
<td>Using test results for summative evaluation purposes</td>
</tr>
<tr>
<td>Identifying underachievers</td>
<td></td>
</tr>
<tr>
<td>Using test results for planning (formative evaluation) purposes</td>
<td></td>
</tr>
<tr>
<td>Using test data to guide remediation</td>
<td></td>
</tr>
<tr>
<td>Recommending counseling or remediation</td>
<td></td>
</tr>
<tr>
<td>Pretesting to determine required instructional emphases</td>
<td></td>
</tr>
</tbody>
</table>
Table 5. (Continued)

Standardized Test Applications

Selection and use of standardized and publisher prepared tests

Norm-referenced vs. criterion-referenced tests

Test norms and interpretation based upon norms

Evaluating tests in terms of reliability and validity

Nontest Evaluation Practices

Class discussion

Observing working habits

Interpersonal relationships

Employing other evaluative devices

Priority topics included in the two columns suggest that teachers and professors have distinctly different desires regarding the course orientation.

This table (Table 5) shows teachers and professors have a common interest in the preparation of exams, but there their commonality ends. Professors want teachers to understand the multitude of ways that test results can be analyzed and information can be extracted and summarized (e.g., group summary statistics) to both best interpret test results and improve test quality. Professors also dwell on standardized testing issues, distinguishing between norm-referenced tests and criterion-referenced tests, as well as dealing with norms, norm interpretation, validity, and reliability.

In contrast, teacher preferences appear to center strictly on classroom instructional decisions. They seem to be saying they want answers to questions such as these: How do I best prepare the test for a given course? How do I administer and score the test? How do I use test information to make specific kinds of decisions? How do I evaluate ongoing classroom actions (e.g., class discussion, working habits, and interpersonal relations)? All are day-to-day issues in the classroom.
In the context of making changes, four of the above-noted undergraduate measurement and evaluation program issues seem most important. First, teachers get relatively little preparation in measurement and evaluation methods. Second, except for the preparation of exams, professors' priorities in measurement instruction do not match either teachers' desired emphases or the way in which teachers apply measurement and evaluation in their classrooms. Third, measurement/evaluation is taught in a context that favors the instruction of fundamental principles, rather than the principles and applications best suited to specific disciplines. Fourth, although the professors appear to have appropriate educational preparation and experience both in educational measurement and in teaching, many measurement professors are not an integral part of the curriculum and instruction program.

NEED FOR CHANGE

For most of us, it comes as no surprise that measurement and evaluation concepts are being taught in a less than totally desirable context. Too little direct instruction is available to the students. Students across all disciplines meet as a group to learn about measurement and evaluation from a professor who is not knowledgeable in all the discipline areas. These students are taught about measurement and evaluation principles in settings where it is difficult to apply directly and practice the measurement and evaluation principles. Such problems are likely to persist regardless of recommended changes.

It seems unlikely that the measurement profession can exert sufficient leverage to increase the amount of time devoted to measurement and evaluation issues, or that the profession can succeed in providing instruction in settings where the students have a common discipline background. Why? Because professors who teach the measurement and evaluation courses are not likely to have a direct say in who takes the course, when the course is taken, or the actual context in which the course is taken.

It could be argued that changes could be made to move instruction into the respective methods courses. Such a move would not necessarily improve the content, and would probably result in a substantial loss in the instructor's measurement and evaluation expertise. Thus, such a move probably would not be a step forward.

Even without changes in program structure, it seems likely that significant changes can occur. Professors appear to have considerable freedom in determining course content. Thus, if persuaded, professors
1. MATCHING MEASUREMENT INSTRUCTION

could take significant individual steps to improve undergraduate teachers’ preparation in measurement and evaluation expertise. For those reasons, the needs and options for change that follow are presented only in the context of changes in the measurement and evaluation course itself.

Presently there appears to be pressure for change in three separate directions: (a) for more direct attention to test development, albeit with different emphases than presently given to the topic; (b) for more attention to non-test assessment; and (c) for greater attention to technological advances, hardware and software applications to facilitate test development, analysis, and so forth (topics that were not even included for rating in the four surveys). The first two options draw much of their impetus from the research findings noted above. The third has impetus primarily because it is new and promising. Additionally, results reported here and elsewhere suggest a strong need for greater attention to design of evaluation and improvement in student feedback mechanisms. Issues surrounding all five of these options are addressed below.

OPTIONS FOR CHANGE

Tests

Both teachers and professors appear to be comfortable with a primary emphasis given to tests. Teachers see tests as valuable and make extensive use of them. Professors appear to be well trained to provide instruction in test development and devote a majority of course time to testing concerns. The major differences here appear with regard to which testing topics should receive emphasis. Professors appear to focus substantial attention on test development, test analysis, and standardized tests. Teachers appear to desire most emphasis on test development and on application of test scores to instructional decisions.

Standardized tests. The apparent difference between the two groups is that teachers want to forgo the preparation in test analysis and standardized tests for additional assistance in application of test results. If the proposed change is viewed from the perspective of teacher testing practices, the change from test analysis and standard test emphases to practical applications appears reasonable. If, however, one views teachers’ desired instruction priorities carefully (Table 4), such a change seems less defensible.

Teacher priorities suggest that teachers want to use classroom test results to make decisions for which classroom tests are not well
suited. For example, teachers want to learn how to use tests (a) to identify gifted pupils or slow learners, (b) to identify underachievers, and (c) to recommend counseling. Such decisions regularly are based upon standardized test results. Thus, a strong argument can be made that if teachers are to make such decisions, then a proper foundation (i.e., study of standardized tests) must be laid.

Many may directly question whether any teachers, let alone beginning teachers, need to or should make decisions about giftedness, retardedness, or counseling matters. In fact, it probably is economically and educationally more sound to leave such determinations to the school psychologist, counselor, or other professional who has substantial training in the use of standardized tests (much more than an introductory course in measurement and evaluation). If this course would be followed, then much of the impetus for emphasis on the practical applications of test results would be removed. Simultaneously, one of the bases for emphasizing standardized tests would be removed as well.

Perhaps the biggest argument for teaching teachers about standardized tests is that students in virtually all schools take standardized tests. Those tests are viewed as an important link between school and home, as indicators both of individual student achievement and of class and school success. Certainly those are important concerns. However, these standardized tests are typically administered at most once a year, and then in only selected grades. Again, perhaps it would be better to depend upon a well-trained individual to coach those teachers who are called upon to use the test results and/or communicate test results to parents.

If there is a willingness to substantially reduce or forgo the emphasis on standardized tests at the undergraduate level, then two things happen. First, the substantial time spent on standardized tests is made available for other emphases. Second, there is much less need to address statistical issues related to the use of standardized tests: reliability coefficients, validity coefficients, standard scores, and the various types of norms.

Teacher-made tests. Both teachers and professors appear to be in such good agreement here that it seems apparent this type of test should receive top billing in the undergraduate course. There are, however, a number of concerns that reside just under the surface.

Teachers and professors profess that tests are good for all instructional decisions, formative as well as summative. Whether tests actually function to serve both formative and summative needs is open to question. Students, for example, view tests as serving
1. MATCHING MEASUREMENT INSTRUCTION

summative purposes. Stake and Easley (1978) note that students view tests as important because their course grades are determined by their test scores. Haertel (1986, p. 10) reached a similar conclusion. He stated, “Student and teacher questionnaire responses confirm that marking and grading is by far the most salient purpose of testing for both teachers and students.”

Can tests function well to provide simultaneously summative and formative evaluation information? Perhaps, but the evidence suggests that they don’t. Haertel (1986, p. 7) found that teachers use tests in a manner consistent with summative evaluation purposes. He writes, “Tests punctuate the flow of instruction, signalling transitions from one unit to the next and bringing closure.” In that context, the purpose of the test is to “tie-off” and close instruction on a topic. That it marks the termination of effort on a selected set of content is evidenced in several ways: The test is preceded by a formal review in class (typically teachers spend nearly a class period in review preparation for the test), the tests are administered in a very formal context (e.g., no use of resource materials and no student interaction), teachers routinely grade their tests, and teachers spend relatively little time reviewing test results with students.

At the point of closure, the posttest review (a formative process?) appears to be deficient in two important respects. First, teachers do not formally analyze tests to look for trends in student understanding or misunderstanding. Thus, the teacher does not go into the review process armed with substantial instructional information. Instead the emphasis is on individual items, the justification of scoring, and piecemeal insights into student understandings or misunderstandings that occur in the review of individual items. Second, once students have received their scores, the payoff has occurred. At that point, for them the test scores represent what they have learned, or failed to learn. They know that learning at that point has low practical payoff because tomorrow they will be responsible for learning a new topic, and what was directly covered by this test will not be directly covered again. Thus, on a need-to-know basis, the content of the test has low priority.

Teachers also argue that a primary purpose of the test is to provide feedback on their instruction. Yet, as previously noted, the teachers surveyed did not take the formal analysis steps that would lead to strong information on whether students reached the desired objectives. Haertel (1986) addressed this same issue in interviews with teachers. He found that although teachers indicated they used
test performance to evaluate their own instruction, only a small percentage could identify any changes in instruction that resulted.

These factors suggest that teachers use teacher-made tests for summative purposes, not formative ones. However, even if tests are downgraded to use as summative tools, they play an important formative role. Teachers teach toward their tests, and students are motivated to study for the tests. In particular, it seems reasonable to believe that students' study will be strongly influenced by the issues and concepts that they expect to be in the test.

Collectively these factors suggest that the primary issue for teacher-made tests is test content. After all, the test content reflects the teacher's instructional objectives, and in a sense directs student study. This suggests that the primary focus on the measurement and evaluation instruction, which relates to tests, should be on test development issues.

Nontest Evaluation

Although the surveys of teacher testing practices have not directly focused on nontest techniques, the issue always emerges. In the surveys described here, for example, teachers first noted the use of their evaluation of students. Then, when asked which topics to emphasize in undergraduate educational measurement and evaluation instruction, they gave nontest evaluation techniques the second highest priority. They want to learn how to evaluate properly using assessment methods other than tests!

Besides teachers' self-perceptions on this issue, findings of measurement professionals support the importance of this topic. Airasian (1984) provides a thoughtful discussion of two general types of nontest assessment, which he calls "Sizing-Up" and Instructional Assessment. In his discussion he outlines the variety of ways teachers routinely access student information and make judgments and decisions that affect instruction and the students' lives.

Haertel (1986), in a study of how teachers choose and use classroom tests, made two important observations about nontest assessment. First, he noted that all teachers interviewed listed affective objectives, but none mentioned any methods for assessment of those objectives. Second, he argued that teachers generally are more balanced in their assessment of students than the students realize. However, he notes that teachers fail to collect, use, and communicate the importance of nontest assessment systematically.

Stiggins and his colleagues at the Northwest Regional Educational Laboratory (NWREL) have conducted the most comprehensive studies
of nontest assessment. Altogether, Stiggins and Conklin (1988) have identified over a dozen assessment techniques used to evaluate achievement, attitudes, and social characteristics of students. Only three of these fit the standard definition of a test. They argue that all methods are equally important, and that each teacher must know how to use properly all of the methods.

These studies, though limited in number, consistently suggest the importance of nontest techniques for classroom evaluation purposes. Teachers attest to their importance. Measurement professionals attest both to their importance and to the lack of appropriate use of such techniques in the classroom. All are strong indicators that nontest evaluation techniques deserve a bigger share of undergraduate measurement and evaluation courses.

Evaluation Design

In a remarkably consistent fashion, the research on teacher classroom-based evaluation shows that such evaluation is a demanding task that requires substantial time and effort. Both the size and complexity of this task point to the need for careful planning to focus and direct the process. Presently, measurement instruction directs little attention to this planning process. The attention provided focuses not on overall design and planning but rather on planning for individual assessment issues. For example, Gronlund’s textbook (1985) provides instruction on the development of instructional objectives, and on the creation of a table of specifications, both in preparation for preparing the test. But nowhere in the text are the issues of general evaluation design directly addressed.

An overall evaluation design needs to be prepared before students walk into the classroom for the first time. That design should prepare (orient) the teacher for a multitude of evaluation tasks including sizing-up, instructional assessment, tests, and more. The size of the planning process and the complexity of classroom evaluation is underscored by those who have looked most closely at the classroom environment.

Stiggins and Conklin (1988) note that the NWREL has identified 12 classroom-level decision-making contexts. Each, they argue, deserves proper assessment prior to determination of a decision. The number of decision contexts alone is clear evidence of the need for careful planning. Twelve decision contexts can beget many more decisions, each decision requiring its own assessment information.

Planning, an evaluation design, organizes the overall perspective on decisions to be made and the contexts within which they will be
made. It forces the evaluator (teacher) to think about matters of timing, identification of assessment processes, selection of instruments, what information needs to be gathered by the teacher, what information must be gathered by the student, what information must be gathered formally, what information can be gathered informally, and a host of other matters.

Failure to carefully plan evaluation of complex situations (and classroom instruction is a complex situation) virtually guarantees misapplication of assessments. That is, it assures that some decisions deserving of formal assessment will be made without assessment information; others will be formally assessed way beyond the needs of the resultant decision. In still others, data will be collected and used that are inappropriate to the decision.

A hallmark of a well-designed evaluation is that assessments are made to evaluate course objectives. Routinely, major gaps can be seen between teachers’ objectives and their assessments. For example, every teacher Haertel (1986) interviewed listed affective outcomes as course objectives; none mentioned any methods that addressed such objectives.

Course grades provide perhaps the best exemplars of evaluation design problems. Grading presents a decision context common to virtually all classrooms. Proper evaluation planning requires that first the rationale for grading be clearly specified in order that information communicated by the grade is clear. Once the rationale and purpose to be served are clear, appropriate data must be gathered to make the grading decisions. Research by Stiggins, Frisbie, and Griswold (1989) strongly suggests that teachers enter into the grading process with neither the rationale nor purpose being clear. They note that teachers routinely gather enough information upon which to base a grade; when they err, it is in the use of too much data. However, because they have not carefully determined the message to be carried by the grade, many teachers incorporate both formative and summative information into the grade. As a result, teachers compute grades from a mixture of assessment information. Some of the information is formally gathered and some is based upon informal impression. Some information reflects achievement; other information reflects nonachievement sources—student attitudes, aptitudes, interests, and citizenship. The net result is reduced validity in grades and less-than-clear communication between teachers and students, as well as between teachers and parents.

Grades are but one example of a multitude of ways that teachers can go wrong through failure to properly design course evaluations.
A standard, almost universal, error made by beginning evaluators is the attempt to gather too much information. The result is that information is gathered, not properly analyzed, and partially used or left to "rot." Time spent in the collection of information that is underutilized could better be spent on other activities. Thus, planning includes the conscious decision to select some decision situations for formal assessment while keeping others informal.

Proper planning also enables the preparation of assessment systems. Haertel (1986, p. 22) argues, "A simple system for recording classroom observations, for example, would make teachers' use of such observations in grading more objective, reliable, and defensible, and would also demonstrate to students that class participation really was considered important." To operate quickly and efficiently, such systems must be thought through and designed beforehand.

Work by Stiggins and Conklin (1988) provides direct evidence that instruction in evaluation design is lacking both in textbooks that teachers use and the courses they teach. Additionally, just how little attention is given to evaluation planning and design is exemplified by Barnes' (1985, p. 47) research. She notes, "... most student teachers equated evaluation with grading or marking papers. Their responses did not convey broader conceptions of evaluation."

Evaluation design must become a part of the preparation that preservice teachers receive. Although the focus on measurement techniques is important, it is not sufficient. We do not expect lessons in how to shoot and use a gun safely to be sufficient to make a person a good hunter. Neither should we expect that attention to tests, checklists, and other evaluative devices will make teachers effective evaluators.

Technology

Recent technological developments are viewed as holding significant promise for improving the capability of teachers to evaluate effectively in the classroom. Ten years ago microcomputers and word processing software did not exist for teacher use. Today, not only are microcomputers and excellent word processing software available for teacher use, but test development programs, item banks, scanners, item analysis programs, and gradebook packages are becoming standard fare.

In the early 1980s, a major question was whether or not item banks and other software would ever be feasible for teacher use. Recent research (Nitko, 1989) suggests that much remains to be done before microcomputer applications can be considered full partners in
the classroom evaluation program. Now, however, the major questions revolve around not whether such technology will be effective, but when it will be of sufficient quality to advocate its use. Soon the questions will be what software and hardware are best and how best to use this new technology. That teachers must be taught to use the new equipment and software is accepted. Whether this technology should be taught as a part of the measurement course or separately (e.g., in a library media course) is a question presently facing most measurement instructors.

Student Feedback

One of the most compelling characteristics of the new microcomputer technology is its capability to display quickly and graphically the results of student assessments. Not only does the computer make it possible to analyze more rapidly student assessment information, it also substantially enhances both the capability to provide feedback and the quality of feedback provided. Germundsen and Glenn (1984) found the ability to provide frequent feedback to students and parents one of the most positive characteristics of a computerized gradebook package they tested.

That issue, communication of information, is perhaps the most overlooked, but most important concern of all. Presently the focus of measurement and evaluation instruction is on the assessment of students to provide information to the teacher. The teacher then is expected to analyze and distribute the information to students.

This channeling of evaluation information through the teacher has two potentially undesirable effects. First, the teacher becomes the gatekeeper of information important to the individual student’s learning. If the teacher decides information is not of import, or if the teacher simply fails to notice or report pertinent information, the student remains unaware.

Second, the process builds a dependency between student and teacher. If the student relies on the teacher to do the evaluative thinking that goes with the learning process, then learning can only progress at the rate dictated by the teacher. Not only is that likely to slow the learning process for the student in the individual course, but the failure to access and use information adequately is likely to carry over into other learning situations as well. We know that students who succeed evaluate effectively. Thus, for the learning process to be most effective, students must not only know what they are to learn, but they must be able to evaluate their personal progress. This requires the development of personal evaluation skills.
These evaluation skills presently are being built into some disciplines, reading and special education for example. Those disciplines appear to provide a structure for evaluation that the student learns to employ for personal instructional advantage. That focus, the planning of evaluation to ensure that students build their own evaluation skills as they learn, is not a part of our measurement instruction. It should be. However, much remains to be learned about how best to employ such practices before that topic becomes an integral part of the undergraduate measurement and evaluation course.

CONCLUSION

The recommendations call for substantial changes in what we teach, if not how we teach, our preservice teacher. If only some of the above recommendations are accepted, the undergraduate course will change substantially. To make these changes requires that some topics be moved out of the undergraduate program altogether. Coverage of other topics will need to be abbreviated.

The argument here is that students must be taught first about the design of evaluation and then about the implementation of evaluation through assessment. If attention is directed first toward the decisions to be made, then evaluation actions can be oriented toward assessment to provide the information necessary to properly make those decisions. This orientation is sure to lead to the choice of instruments and assessment activities to serve the desired needs. Attention to individual instruments, and the trade-offs in using different instruments and strategies, then comes naturally.

REFERENCES

Airasian, P. (1984, November). Classroom Assessment and Educational Improvement. Keynote address at a conference titled Classroom Assessment: A Key to Educational Excellence, Northwest Regional Educational Laboratory, Portland, OR.


The current state of teacher training in assessment has been thoroughly documented in previous chapters. The resulting picture is one of neglected and irrelevant training in an arena of professional activity that forms the basis of sound instruction. The decisions teachers must make cannot be made well without sound achievement data. The decisions students make about themselves cannot be made well if those students do not receive sound information on their achievement. The decisions made by those in leadership positions cannot be made well without the sound achievement information that comes from sound assessment. Obviously, high-quality assessment is crucial to the development and presentation of sound educational programs. And yet, we see before us a picture of professional development for educators that is almost completely devoid of assessment training.

Our recently completed, decade-long task analysis of classroom assessment has revealed that teachers typically spend a third of their professional time or more involved in assessment-related activities. They use assessments almost continuously to inform a wide variety of decisions and to serve other purposes that directly influence the quality of the learning experiences provided to students (Stiggins & Conklin, 1992). If school improvement efforts are to succeed, they
must include a component that teaches teachers how to use this massive amount of in-class assessment time productively.

In this chapter, I plan to add a few brief insights from the Pacific Northwest to the emerging portrait of teacher training in assessment. Our picture is not different from those already described. It is a picture of neglect. Very few teachers in our region are offered the opportunity to participate in relevant classroom assessment training.

Next, I will discuss some of the possible reasons for this unfortunate neglect. Why has so critical an area of professional competence been given so little attention in teacher preparation for so long?

The third issue I will address is that of the mismatch between (a) what teachers need to know about assessment in order effectively to manage classroom assessment environments and (b) what they are taught about assessment during their professional preparation, if they are offered any training at all. Our analysis of the task demands of classroom assessment has yielded a clear framework of classroom assessment competencies for teachers. I will compare the assessment training currently offered to these essential competencies.

Then to conclude, I will discuss the actions we need to take to eliminate the mismatch. Given the neglect of training and the irrelevance of training when offered, what do we do to provide relevant, helpful training to teachers? How do we revise training priorities to make this training attractive to teachers? And how do we let policymakers know that resources must be allocated to provide this previously neglected training?

Assessment Training in the Northwest

In our investigation of the current status of teacher training in assessment in the Pacific Northwest, we examined teacher certification regulations to determine requirements for assessment training, and we analyzed the assessment courses offered in the major teacher training programs of the six-state region, which includes Alaska, Hawaii, Idaho, Montana, Oregon, and Washington (Stiggins & Conklin, 1988). Within these states, we reviewed 27 undergraduate and graduate teacher training programs across 14 teacher training institutions. These programs produce 75% of all of the teachers graduated annually in the region. Our analysis asked whether assessment courses were offered, whether they were required for graduation, and what content is covered in these courses.

Only one of the six states (Oregon) explicitly requires assessment training for certification. All others require graduation from an accredited teacher training program. In addition, many require
candidates for licensure to attain a certain minimum score on the National Teacher Examination (NTE).

Our analysis of a sample NTE reveals that only 11 of the 339 test items address assessment issues, and only 4 of these deal with assessment issues that are directly relevant to classroom assessment for teachers.

Our analysis of the teacher training curriculum reveals that only 13 of the 27 programs currently offer an assessment course and only six programs required completion of that course for graduation. From this, we concluded that the vast majority of teachers currently practicing in the region probably received no assessment training whatever as part of their professional preparation. Further, our analysis of the content of that training reveals that, even when training is offered, it fails to match the training needs of those who must develop and use assessments on a day-to-day basis in the classroom. Before discussing this mismatch, however, I want to explore some of the possible reasons why assessment training is so totally neglected in so many programs.

Reasons for Neglect

We have been able to identify at least five possible reasons why assessment training is so frequently excluded from the teacher training curriculum. In fact, the true origin of this problem probably resides in some combination of these and we may never be able to disentangle the contribution of each. But each possible reason implies some actions we can take to remedy the situation. So it is in our best interest to strive to understand each.

One possible reason for our neglect of assessment training might be our tendency to focus on process rather than outcomes in the management of education. For example, high school graduation decisions traditionally have been based on the completion of certain credit hours rather than the attainment of certain outcomes. In this case, the assumption is made that, if the credits are completed (the process variable), the outcomes will take care of themselves. For another example, schools often define the teacher’s job in process terms, such as when teachers are evaluated in terms of whether they complete the textbook in the allotted time or not. This definition of good teaching assumes that covering the material at a certain rate (the process variable) will produce maximum learning (the desired outcome). Yet another example can be found in our procedures for accrediting schools. The accreditation decision traditionally has rested on the
evaluation of such factors as faculty credentials, student/teacher ratios, adequacy of facilities, etc. Again, the assumption is made that proper process leads naturally to desired outcomes. In an environment where process-oriented evaluations of students, teachers, and programs rule the day, training in the assessment of outcomes may not be regarded as central to the evaluation task and therefore may not be included in professional preparation programs.

Another possible reason for the absence of assessment training in teacher training programs may be the fact that these courses often have a reputation as being somewhat more academically demanding than typical education courses. In my teaching experience, I see many teachers put required assessment courses off to the very end of graduate programs due to their anxiety about such courses. Over the years, perhaps these tougher academic standards have made such courses unpopular with students and other faculty, and thus have resulted in their elimination from programs.

A third, more subtle reason for the neglect of this kind of training may be the fact that the systematic assessment of outcomes may be seen as being too risky by school personnel. If schools are very clear about their achievement targets, and are clear and public about their assessments of those outcomes, there is always the chance that someone in the community will disagree either with the target or the assessment. Or there is always the possibility that students will be found to have learned already what we had planned to teach them before we have a chance to teach them. Or further, there is the danger that either we and/or the public might discover after instruction that students failed to learn to hit the target. Under any of these circumstances, time and energy will need to be expended with the hassles of defending our priorities, reorganizing our efforts, individualizing instruction, and/or revising programs. In this kind of environment, educators may regard it as safer and easier to keep the achievement targets vague and to keep our assessments broad and out of focus. Further, we may regard it as safer simply to remain naive about key assessment issues. Systematic assessment training may not be a high priority for educators concerned about public review or the possibility of change.

Yet another possible explanation for the neglect of assessment training may be the assumption on the part of educators that the quality of assessments in the classroom is assured from outside the classroom; that is, quality assessment is assured by means beyond the control of the teacher. For instance, textbooks often are accompanied these days by their own quizzes, unit tests, and even computerized test item
banks for teachers. We may conclude, therefore, that it is unnecessary for teachers to know how to develop their own assessments. Besides, even if the text-embedded or teacher-developed assessments don’t keep the standards of achievement or test quality as high as we would like, we may find solace in our belief that we can count on those very high-quality standardized tests to bolster our standards of excellence. If we believe these things to be true (whether they are or not—often they are not), we are less likely to value assessment training for teachers.

Without doubt, each of these four factors has contributed in some way to the current state of neglect in the assessment training of our teachers. But I believe the major cause rests not in our process orientation to evaluation, or the fact that testing courses are too tough, or the fact that systematic, public assessment is too risky, or even in our false confidence that we have teacher-proof assessments in place. Rather, I believe the explanation resides in the historical irrelevance of the assessment training we have offered. The concepts covered, the assessment strategies taught, and the assessment quality control procedures advocated in assessment courses traditionally have failed to reflect any whatever sensitivity to the realities of the classroom. I will document the exact nature of this failing in precise detail in the next section. In the meantime, suffice it to say that, in an environment where credit hours for teacher training have always been restricted and currently are declining, what teacher training institution is likely to waste valuable credits on coursework that bears little resemblance to the realities of teaching in the classroom?

Training Versus the Realities of the Classroom

Our research analysis of the task demands of classroom assessment has suggested six specific dimensions of classroom assessment environments that teachers must manage effectively if they are to integrate sound assessment into affective instruction. Each dimension suggests a set of assessment competencies teachers must master if they are to reach this goal. Those dimensions and their associated competencies hold that teachers must understand the:

- full range of possible uses of classroom assessment
- achievement targets they hold as expectations for students and how those targets translate into assessments
- qualities of a sound assessment
- full range of assessment tools at their disposal
- critical interpersonal dimensions of classroom assessment
keys to formulating and delivering feedback on assessment results

Let's analyze each of these, comparing what teachers need to know about each with what they are taught in the few assessment courses we found in the teacher training curriculum of the Northwest.

**Classroom uses of assessment.** Our analysis suggests that teachers use assessments in their classrooms to serve at least three different categories of purposes. First, they use assessment results to inform decisions. They diagnose student needs, select students for special services, group students for instruction, assign grades, etc. Second, they use assessments as teaching tools, such as by using them to communicate achievement expectations to students, using assignments both as practice and as assessments of achievement, involving students in self and peer evaluation to help them become better performers, using practice tests, etc. And third, they use assessments as a classroom management or behavior control mechanism to keep students in line. Assessment is the major power tool of the classroom environment and teachers control the switch.

If they are to use assessments in all of these contexts in a fair and effective manner, teachers must understand how each use relates to quality instruction, what role assessment can play in each use, and how the situational variables associated with each use impacts the meaning of a quality assessment.

Our analysis of currently available teacher training courses in assessment reveals treatment of only the first category of purposes, those related to decision making. And even in this case, the coverage is superficial, dealing only with the distinction between criterion-referenced and norm-referenced tests and their relationship to various decisions in the classroom and at higher levels of the education organizational hierarchy. We found no treatment of assessment as a teaching tool and virtually no comment on or guidelines for the use of assessment as a behavior management tool—both obviously critical aspects of effective classroom assessment. And we found no treatment of issues related to changes in the meaning of assessment quality as assessment purpose varies.

**Achievement targets in assessment terms.** One of the basic tenets of sound assessment in any context is that the assessor possess (a) a clear and highly differentiated vision or understanding of the achievement target to be attained by students, and (b) a thorough understanding of the full range of assessment alternatives available to assess the target of interest. It is impossible, for example, for a teacher to assess a
student’s level of writing proficiency if that teacher does not have in mind a clear vision of what it means to write well—a clear sense of the attributes of good writing. The same is true of the assessment of good reading, thinking, speaking, mathing, sciencing, etc. Certainly it is not the responsibility of the assessment course instructor to teach teachers to have these visions of desired outcomes. That is the responsibility of the content area instructors. However, it is the responsibility of the assessment instructor to provide guidelines for the translation of the various targets into proper assessment methods.

Our analysis of the task demands of the classroom reveals that teachers expect their students to aim for, and must assess, at least five different kinds of achievement targets: First, there is almost always some specific substantive subject matter knowledge to be mastered. In addition, teachers often want students to be able to demonstrate higher order thinking or problem-solving skills using that knowledge. Third, most teachers hold expectations that students will be able to demonstrate certain specific achievement-related behaviors. Fourth, many teachers want their students to be able to create certain achievement-related products that possess certain attributes. And finally, teachers often hope students will attain certain affective goals.

Teachers need to understand how all of the various types of targets translate into assessments. They need to complete assessment training with sufficient practical know-how to be able to align assessments with all of the various types of valued achievement targets.

Our analysis of the achievement targets addressed in the assessment courses we studied reveals the treatment of only two of the four kinds of achievement targets: knowledge and higher order thinking. Strategies are presented for assessing these valued outcomes through the use of paper-and-pencil assessment tools. This is important training that will be of great value in most classrooms. But it is by no means sufficient.

First of all, the definition of higher order thinking advanced in assessment courses almost universally is the definition presented in the Bloom taxonomy of cognitive levels (Bloom, 1956). This represents only one of many such definitions available to teachers. They need to become aware of the full range of alternative conceptualizations at their disposal. Many of the others are far easier than Bloom for teachers and students to deal with. The Quellmalz (1985) taxonomy represents one excellent example. Thinking skills targets need a much broader treatment in assessment training.

Second and most importantly, assessment training needs to address the other three kinds of achievement targets most often com-
pletely ignored in the courses we studied: achievement-related behaviors and products, and affective outcomes. These do translate into systematic classroom assessments and teachers need to know how to do so. We must strive to disimbue ourselves and teachers of the notion that all (or even most) of the achievement outcomes we value for our students can be translated into objective test item formats. They cannot. Teachers need to know how to translate all their targets into assessment terms. Currently available training does not offer this.

The qualities of sound assessment. We know that the definition of a high-quality assessment varies as the assessment context changes. Therefore, it is not possible to give teachers a specific formula for quality to apply in a rote manner in the classroom. However, we also know that there are a few general quality-control guidelines that teachers must understand, so they can adapt them to the various assessment contexts they face on a day-to-day basis. For example, they must know that quality assessments:

- arise out of a clear and specific target and reflect that target in their assessment methodology
- control for various sources of extraneous interference that can cause us to mismeasure achievement, such as attributes of the student, the assessment process, and/or the assessment environment that are unrelated to student achievement but that influence test results
- sample student performance in a manner that is representative of the performance domain and is sufficiently large to justify our conclusions, yet is economical in that it does not produce more information than we need to the purpose
- provides the users with information in a form they understand and that fits the purpose

Each of these attributes of sound assessment implies a different set of potential sources of mismeasurement. Teachers need to know how to avoid all of these pitfalls. They need to know how to identify a mismatch between a target and an assessment method and how to fix it. They especially need to know all of the various sources of extraneous interference that can pop up both with objective and subjective assessment and how to prevent the problems from occurring. They need to know about potential sampling problems and how to avoid them. And they need to know how purpose and assessment method link up and how to evaluate whether they or other users (e.g., students) truly understand the information resulting from an assess-
ment. There are certain very practical procedural steps teachers need to understand to promote sound classroom assessment.

We did not see these addressed in the courses we studied. Rather, we saw issues of quality control in assessment being addressed from a completely different perspective. That treatment of quality focuses on (a) the definitions of various types of validity and reliability, and (b) the statistical estimation of the validity and reliability of objective tests. Neither of these treatments has practical relevance to teachers in classrooms. They do not help teachers produce and use quality assessments. Far greater attention must be given to eliminating sources of measurement error.

Assessment tools. Teachers use at least three forms of assessment in tracking student achievement on a day-to-day basis in the classroom. First, they rely on paper-and-pencil assessment instruments, including teacher-developed and text-embedded tests and quizzes, assignments, standardized tests, and questionnaires. In addition, they rely on observations of and professional judgments about achievement-related behaviors and products. And third, they rely on direct personal communication with students to find out what they are learning, such as through instructional questions, interviews, casual conversations, discussions with others, and intuitions and feelings about students and their needs.

Each of these methods has strengths and weaknesses when used in various contexts. Each matches up well with some achievement targets and not others. Each carries with it a unique set of problems and pitfalls to be avoided in its design and use. Teachers need to understand these things about each set of tools.

The courses we studied covered these topics for only one set of assessment tools: paper-and-pencil instruments. And this coverage was limited to teacher-developed and text-embedded tests and quizzes and standardized tests. Assignments as assessments were ignored, as was the development or use of questionnaires. Further, the vast majority of courses paid little attention to the use of observation and judgment as assessment, and all courses virtually ignored personal communication as a mode of assessment in the classroom.

Each of these kinds of assessment can be done well or poorly. Each carries with it certain unique rules of evidence for sound use. The fact that teachers need to know these things seems to have been completely missed by course designers.

Interpersonal dimensions of classroom assessment. Classroom assessment environments are complex interpersonal places. Assessment is
virtually never a detached, scientific, objective laboratory act of dipping the dipstick to test the level of learning. Rather, it is virtually always an interpersonal act with personal antecedents and personal consequences. Experienced teachers know this perhaps better than anyone. But they often are unaware of the implications of this fact for the assessment methods they use. They often overlook the specific impact of assessments on their students as people.

There are a number of interpersonal facets of classroom assessment that need to be covered in depth in training. These include the facts that:

- students are key contributors to the classroom assessment process and environment, because they:
  * come from vastly differing home cultures, some of which directly impact the assessment of their achievement
  * hold expectations of themselves derived from teachers' classroom assessments of them
  * are consumers of assessment information as self-assessors and crucial decision makers
  * maintain a sense of control over their own academic well-being based on their own assessments of the achievability of achievement targets
  * are peer assessors, judging each other and forming relationships based in part on academic performance in the classroom
  * differ widely in their understanding of the implicit curriculum and what it takes to look like a high achiever
  * differ widely in temperament, assessment anxiety, feedback needs, and motivation to learn and be assessed during the learning process

- teachers are key contributors to the interpersonal assessment environment of the classroom in that they:
  * hold widely differing expectations of students
  * have differing personal reactions to students as people
  * hold all of the power of control over classroom life in their power to assess and evaluate
  * differ widely in temperament, sensitivity, and motivation to teach and assess learning

Out of these important dimensions of classroom assessment environments there arises a set of competencies teachers must master if they are to treat students in a sensitive and equitable manner from an
assessment point of view. And yet, nowhere in the courses we studied were we able to find any evidence of the treatment of these crucial issues.

**Feedback on assessment results.** Teachers continuously formulate and deliver feedback on assessment results. This too is a critical aspect of the academic and interpersonal environment of the classroom. Although all forms of feedback are important, one very prominent form exerts greater influence than the others and therefore deserves special attention. That form is report card grades.

With respect to grades and grading, teachers carry out effective practices when they communicate those practices to students in advance, so students know what is expected; factor various student characteristics into the grade that belong there (e.g., achievement) and leave out all else (e.g., attendance, personality, attitude); use sound achievement data as the basis for grades; keep thorough, appropriate records; and combine data carefully over time and set appropriate cutoff scores to determine report card grades.

With respect to the other forms of feedback teachers use, such as oral communication, nonverbal communication, written comments, performance ratings, and test scores, teachers carry out sound practices when they focus feedback on clear expectations, time feedback to ensure student attention, and check for understanding of feedback. Teachers need to learn these things somewhere in their professional preparation.

Yet again, as with the interpersonal dimensions of classroom assessment environments, we found the arena of feedback on assessment results to be completely neglected in the courses we studied.

**Summary.** As a result of years of study, we know what teachers need to learn about assessment to function effectively in the classroom. Our belief glimpse into the assessment training of teachers in the Northwest reveals that they are not being taught what they need to know. Two of the six key competency arenas (interpersonal aspects and feedback) are being completely ignored, while the others (assessment purposes, achievement targets, qualities of sound assessment, and assessment tools) are being treated so narrowly and with such lack of depth as to render currently available training almost useless to teachers.

**Changing Direction**

Inadequate classroom assessment has direct implications for all. Students who succeed in hitting the target but who fail anyway due
to inept assessment lose their sense of control over their own academic well-being. Teachers face the prospect of less-than-effective or inefficient instruction and, in addition, feel a growing sense of alienation from testing—one of the keys to their success. The public continues to view schools through a filter of assessment illiteracy that allows them to continue to assume naively that all or most of the achievement outcomes we value for our students can be assessed via published norm-referenced standardized tests. School improvement efforts continue to have less impact than they need to have, because all of the effort devoted to attaining better outcomes is expended by those unable to assess whether those outcomes have been attained as a result of program improvements. This list of implications could go on for pages. Sound, relevant assessment training for teachers (and other educators) is an absolute must.

How then shall we reach this goal? I have several suggestions for immediate action.

First, we must deal with each of the five potential reasons for neglect of assessment training cited earlier. And to a very real extent we are beginning to do so. We must reorient from process-based to outcome-based evaluations of students, teachers, and programs. We are starting to do this, although these efforts are just beginning. High school proficiency assessments are becoming more common. Teachers are being held accountable for outcomes. And accrediting agencies also are examining outcome data. As these trends grow, high-quality, relevant, helpful assessment training will become a higher priority for all.

If assessment courses have been more academically demanding and students have had difficulty hitting the achievement targets designated by assessment instructors, we must analyze both the targets and the teaching methods used in these courses. Clearly, as I described in the previous section, the achievement targets for these courses have not been appropriate. Although we cannot judge the quality of instruction based on our study, we do know that if instructors become good teachers, modeling these methods for teachers, and evaluating the performance of their students using the proper methods, the probability will increase that future teachers will meet the demanding standards of assessment training.

If school personnel are uneasy about the dangers of being clear about achievement targets, and systematic and public about the assessment of those targets, then a higher level of assessment literacy on their part can only help. It will help because assessment training will give educators the tools and wisdom they need to be sure (a) the
public understands the full range of achievement outcomes we expect of our students (the public currently does not understand this!), (b) to develop and use the full range of assessment methods needed adequately to represent student attainment of those outcomes (norm-referenced standardized tests do not do this!), and (c) to plan instruction that directly treats valued achievement targets, thus greatly increasing the probability of student success at all levels of the achievement continuum (including advanced, average, and perpetually failing students!). None of these goals can be achieved by an education community that is essentially illiterate with respect to assessment issues. In fact, the risk of unfavorable public review is far greater if we remain uniformed in this critical arena.

If we believe teachers need not understand assessment because someone else already has taken care to assure quality classroom assessment, we need only examine the quality of many text-embedded tests and quizzes. Many of these are developed in the complete absence of quality control standards. If we believe standards of educational excellence are maintained by standardized tests, we need only think about (a) the extent to which these tests cover the full range of valued outcomes and (b) the fact that teachers make decisions about how to interact with their students at the rate of one every few minutes, whereas standardized tests happen only once every year or so. The standards of assessment quality and educational excellence can only be maintained if each teacher in every classroom is the best assessor he or she can possibly be.

Finally, if we currently neglect assessment training because that training historically has been irrelevant, we need to make the training relevant and helpful. The entire premise of this chapter is that we know how to do this. We need only make it a priority and allocate resources to make it happen.

Even as we deal with the various causes for neglect, there are other specific actions we can take:

1. Place a priority on in-service training. We are a national faculty that graduated from professional preparation programs that included no such training.
2. Design public relations programs to convince teachers and other educators that systematic classroom assessment can make their assessments (and therefore their teaching) faster, easier, and better, in that order. That is, sell assessment as the time and energy saver that it can be.
3. Separate assessment training audiences. The training needs of teachers are unique. They are not the same as guidance
counselors, psychologists, Ph.D. candidates in testing, or even educational administrators. Teachers should be trained separately.

4. All educational administrators should be required to complete training in classroom assessment and large-scale assessment. Only then can they (a) support the efforts of their teachers and (b) communicate with the public about school attainment of intended outcomes.

5. Assessment course instructors must understand the realities of life in classrooms. All who have not spent time in public school classrooms, or have not been there recently, should go to observe and teach there. This will reveal to them the complexity of the assessment task demands teachers face every day.

6. Through this in-class experience, assessment instructors also can learn from good teachers the basic principles of good teaching. These principles can be applied to the development and presentation of sound assessment instruction also.

In short, assessment training has a terrible reputation to overcome. It is regarded as irrelevant, technically complex, academically demanding, and a waste of valuable credit hours. Many teachers have had bad experiences with this training. Unfortunately, this reputation is deserved.

The time has come to change both the image and the reality of assessment training for teachers and other educators. This is partly a problem in public relations—a problem in salesmanship. But before the new product—relevant, helpful assessment training—can be sold effectively, it must be developed. We have all of the necessary ingredients in hand. We need only assemble them properly and put them in place everywhere students are assessed and evaluated.

REFERENCES


THE DEVELOPMENT
OF STANDARDS FOR TEACHER
COMPETENCE IN EDUCATIONAL
ASSESSMENT OF STUDENTS

James R. Sanders and Suzanne R. Vogel

Western Michigan University

There has been a long history of concern about the quality of student assessments and their use by educators, and rightly so. Test scores, grades, informal measurements, and other forms of assessment typically have been weighted heavily in decisions about students, programs, and policies. Malpractice in student assessment can have detrimental and irreversible consequences affecting human lives and school programs. Assessment is defined here as the process of obtaining information that is used to make educational decisions about students; to give feedback to students about their progress, strengths, and weaknesses; to judge instructional effectiveness and curricular adequacy; and to inform policy.

The National Council on Measurement in Education (NCME) studied the feasibility of credentialing measurement experts in education, and concluded that because the practice of measurement and assessment is so pervasive in education and takes on so many different forms, it would be much too costly to develop credentialing procedures for every type of assessment practice (Sanders, 1987). As
an alternative, the NCME undertook the development of standards of assessment competence for major practitioner roles in education: classroom teachers, school administrators, counselors, testing directors, curriculum specialists, and others. In 1987 the NCME invited three other professional associations to collaborate on the development of standards for classroom teachers, the largest practitioner group and the one that uses student assessments most frequently. Similar collaborative projects, focused on other educational practitioners, are expected to follow.

The collaborators on the teacher standards were three associations directly involved in the preparation and professional development of classroom teachers: the American Association of Colleges for Teacher Education (AACTE), the American Federation of Teachers (AFT), and the National Education Association (NEA). The report of this project was published in 1990.

In the remainder of this chapter we will review selected literature on teacher preparation in student assessment: (a) how classroom teachers use measurement and student assessments in the classroom, (b) what experts have said teachers need to know about measurement and student assessment, and (c) the status of training prospective teachers in student assessment. We will then describe the standards developed by the four collaborating associations, and conclude with a brief discussion of work that still needs to be done to improve the quality of student assessments and their use in education.

LITERATURE ON TEACHER PREPARATION IN STUDENT ASSESSMENT

The need for developing standards to guide teachers’ professional preparation and in-service training in assessment was recognized as far back as 1912 (Starch & Elliot, 1912), and has been building since 1967 when Samuel Mayo presented his report, Pre-service Preparation of Teachers in Educational Measurement, and David Goslin wrote Teachers and Testing. The importance of assessment competence for teaching was highlighted by Rudman, Kelly, Wanous, Mehrens, Clark, and Porter (1980), who described the necessity for teachers to use a variety of assessment methods in order to make appropriate decisions about student grading, grouping, placement, and instruction. The ability to use information properly when making important student, instructional, or curricular decisions is an integral part of professional teaching practice. Research has consistently revealed, however, that the preparation of teachers at most universities in the area of assessment is either inadequate or totally absent (Noll, 1955; Roeder, 1972,
How Classroom Teachers Use Measurement and Student Assessment in the Classroom

Gullickson (1985) conducted a survey of 295 South Dakota teachers to determine the relationship, if any, between 11 student evaluation techniques, grade level, and curriculum area. His study showed that the most highly rated techniques across all grade levels and curricula were objective teacher-made tests, discussion, and papers/notebooks.

According to Gullickson’s 1985 report, elementary teachers tend to rely on several evaluation techniques of pupil progress. “Class discussion, evaluation of student papers, and evaluation of student behavior all are seen to hold a higher priority than tests” (p. 99). The elementary teachers do tend to give more credence to the results of standardized objective tests than do junior and senior high teachers.

According to the results of Gullickson’s survey, secondary (junior and senior high) teachers tend to rely on fewer evaluation techniques, with teacher-made objective tests being the method of choice. Secondary teachers reported that they use essay tests much more frequently than do elementary teachers.

In a more extensive survey of classroom teachers in South Dakota (336 respondents), Gullickson investigated purposes for testing, frequency of testing, sources of test items, and preferred methods of measurement. The findings of this study are consistent with the previous study in that generally teachers rated teacher-made objective tests most highly. Secondary teachers again placed significant emphasis on essay tests. These evaluation techniques were followed in order by standardized objective tests and oral quizzes (Gullickson, 1982).

Teachers reported using tests frequently, with 95% indicating weekly use of tests. Gullickson’s study indicated that teachers spend a great deal of time in test-related activities, with the estimated average time spent in such activities being 190 minutes per teacher-made objective test. Assuming that teacher-made objective tests are administered on a weekly basis, this translates into about one-half teacher day per week spent on test-related work.
Teachers reported that they author their own test items 93% of the time, use items from textbook publishers 60% of the time, and use other published test items 23% of the time. When asked to indicate the types of items normally used on their tests, 92% indicated short answer/completion, 77% matching, 76% multiple choice, 67% true/false, and 58% essay, with 31% of elementary teachers using essay as opposed to 69% of secondary teachers. Teachers indicated that about 75% of their course content is covered by their teacher-made objective tests.

The great majority of teachers in Gullickson's study indicated that the following test administration conditions are the norm:

- Students may not interact.
- Students may not use resource materials.
- Students may not use calculators, except in senior high science courses where 40% of teachers allow their use.
- Tests are not speeded.

Sixty-four percent of the teachers reported that they do not use separate answer sheets.

The overwhelming majority of teachers (97%) reported that they always or usually score their own tests. Only 55% report that they always or usually provide written comments on tests. The vast majority of teachers (90%) use total score as the only means of test analysis. Forty-two percent of the teachers use score range. Mean, median, and standard deviation are used by relatively few teachers in test analysis. Roughly one third of the teachers report analysis of item difficulties and test reliability.

Teachers reported that they generally grade (95-97%) their own tests, and 94% return tests promptly to students (within 2 days). Relatively little time is spend during class time for posttest review (Gullickson, 1982).

Gullickson’s results confirm many of the findings of Fleming and Chambers (1983), who conducted systematic analyses of teacher-made tests in the Cleveland, Ohio Public Schools in response to a federal court order for desegregation. The authors made the following observations about how Cleveland teachers test:

First, teachers use short-answer questions most frequently in their test making. Second, teachers, even English teachers, generally avoid essay questions, which represent slightly more than one percent of all test items reviewed. Third, teachers use more matching items than multiple-choice or true-false items. Fourth, teachers devise more test questions to sample knowledge of facts than of any other behavioral categories studied. Fifth, when categories related
to knowledge of terms, knowledge of facts, and knowledge of rules and principles are combined, almost 80% of the test questions reviewed focused on these areas. Sixth, teachers develop few questions to test behaviors that can be classified as ability to make applications. Seventh, comparison across school levels shows that junior high school teachers use more questions to tap knowledge of terms, knowledge of facts, and knowledge of rules and principles than elementary or senior high school teachers do. Almost 94% of their questions address knowledge categories, contrasted with 69% of the elementary school teachers’ questions. Finally, at all grade levels, teacher-made mathematics and science tests reflect a diversity of behavioral categories, since they typically feature questions in all six behavioral categories. (p. 32)

Fleming and Chambers (1993) found that teachers generally used one-page tests that were usually neat in appearance, but because of poor quality reproduction were sometimes difficult to read. Teacher-made tests often did not contain clear directions, and were found to have errors in punctuation and spelling nearly 20% of the time. Teacher-made tests lacked indication of point values for test items in most cases, which suggests to the authors that “teachers may not be visualizing their tests as a means for quantifying students’ performance as a measure of students’ learning” (p. 36).

The Cleveland study indicated some problems with item construction. For example, multiple-choice item stems might be only one or two words; short answer/completion items might be unclear; multiple-choice items might have more than one defensible correct response. The authors concluded that their review of teacher-made tests “seems to indicate that training programs addressing item construction and tests as measurement of student learning are desirable” (p. 37).

What Experts Have Said Teachers Need to Know About Measurement and Student Assessment

Measurement specialists and educators have long voiced their views about what teachers need to know in the area of measurement. In 1964 Mayo conducted an extensive survey of teachers, principals and superintendents, college and university professors, and testing and research specialists. His purpose was to identify an ideal list of competencies for beginning teachers in the area of educational measurement.

Mayo’s (1967) survey results seemed to indicate that many respondents placed equal emphasis on teacher knowledge of standardized testing and classroom or teacher-made tests. Gullickson’s two
studies (1982, 1985) and the findings of Fleming and Chambers (1983) seemed to indicate that teachers rely most heavily on teacher-made tests for student evaluation and classroom instructional feedback. It would be very easy to conclude that measurement instruction for teachers should be concentrated on such areas as test construction, grading, item analysis, and establishment of reliability and validity. Fleming (1979) spoke to the issue of real-world classroom measurement: the routine use of teacher-made tests versus standardized tests to measure students’ learning. Although she agreed that standardized tests are not always indicative of material taught in the classroom, and that teacher-made tests may be preferable, she voiced clear concerns about the quality of teacher-made tests. She contended that the children in the classroom receive much more information about their learning from the teacher-made tests they routinely take than from standardized test results that usually do not affect student grades, and the results of which may never even be reported directly to the students. “Certainly the failure message is communicated much more frequently from the classroom test than the standardized test” (p. 5). Because of the possibility that failure messages are communicated to students due to faulty measurement instruments, Fleming proposed the following as classroom measurement needs in the 1980s, requiring the support of school districts:

1. There should be renewed efforts to improve preservice and inservice training in evaluation of instruction. Evaluation should be emphasized as a critical step within the teaching cycle.
2. There is a need for more effective and comprehensive training materials in educational evaluation.
3. There is a need to improve the operation of their district-wide measurement systems as a support to improvement of classroom measurement processes.

Additionally, Fleming identified the following needs in the area of instrumentation:

1. There is a need for improved teacher-made classroom tests at every level.
2. There is a need for assessment procedures which may be utilized within the emerging “new” models for teaching.
3. There is a need for improved procedures for measurement of writing.
4. There is a need for development of language assessment instruments for the support of bilingual programs in the schools.
5. There is a need to develop naturalistic methodology which has application to classroom assessment problems and which has utility for classroom teachers.
6. There is a need for developing options in criterion referenced measurement for the classroom teacher. (pp. 1-20)

The case presented by Fleming and Chambers and by Gullickson for concentration on teacher-made measurement is indeed strong. However, teachers, particularly elementary teachers, report using results of standardized tests (Gullickson, 1985). Rudman et al. (1980) provided some additional insights as to the reasons for emphasizing teacher knowledge about standardized tests. They indicated that teachers make critical decisions regarding student placement and programming early in the school year, and require information within the first 3 or 4 weeks of school in order to make such decisions. Many of these decisions are affected by results of standardized test scores available in the students’ files, as well as by teacher observations and intuition. Additionally, teachers may be responsible for the interpretation of standardized test scores to parents at parent-teacher conferences. Rudman et al. (1980) concluded that teachers need a variety of information sources in order to make appropriate decisions about grouping, placement, and instruction. Assessment and instruction should be incorporated in the classroom, and classroom teachers need the knowledge and skills to make this possible.

Other authors have attempted to identify measurement competencies needed by classroom teachers in broader terms. Robert Ebel (1962) developed the following principles of measurement for educational achievement:

1. The measurement of educational achievement is essential to effective education.
2. An educational test is not more or less than a device for facilitating, extending, and refining a teacher’s observations of student achievement.
3. Every important outcome of education can be measured.
4. The most important educational achievement is command of useful knowledge.
5. Written tests are well suited to measure the student’s command of useful knowledge.
6. The classroom teacher should prepare most of the tests used to measure educational achievement in the classroom.
7. To measure achievement effectively the classroom teacher must be (a) a master of the knowledge or skill to be tested and (b) a master of the practical arts of testing.
8. The quality of a classroom test depends on (a) the relevance of the tasks included in it, (b) the representativeness of its sampling of all aspects of instruction, and (c) the reliability of the scores it yields.
9. The more variable the scores from a test designed to have a certain maximum possible score, the higher the expected reliability of those scores.
10. The reliability of a test can be increased by increasing the number of questions (or independent points to be scored) and by sharpening the power of individual questions to discriminate between students of high and low achievement. (pp. 21-26)

Ebel’s principles reflect an underlying agreement among the experts that measurement must be incorporated routinely into the instructional process. Farr and Griffin (1973) indicated that teachers need to be shown the close relationship between measurement and instructional decision making. They asserted it is perhaps too often the case that measurement is dealt with in the preservice education of teachers as an entity unto itself, with the result that “the basic principle underlying the discussion of what teachers need to know about measurement is that measurement should serve a purpose” (p. 19) is neglected. They developed the following “Outline of Measurement Concepts and Skills Needed by Classroom Teachers”:

**Listing Instructional Decisions**
A. For which decisions can information be collected?
B. Which decisions require continuous information feedback and which require only periodic feedback?
C. Are the decisions consistent (valid) with a stated definition of the skills and behaviors to be taught?

**Developing Decision Alternatives and Determining Information Needs**
A. What are the measurable differences between alternatives?
B. What criterion [sic] are used to determine the feasibility of particular alternatives?

**Collecting Information**
A. How can information be collected validly and reliably?
B. What procedures are there for collecting information congruently with instruction?
C. What are the strengths and weaknesses of various data collecting procedures?
D. How can collected information be related to decision making?
E. How can teacher observations be made more valid and reliable?
F. How should teacher assessments be constructed? (p. 27)

Farr and Griffin believed this outline could serve as a guide in the development of teacher competencies in measurement that directly relate to the classroom behaviors of teachers.
3. DEVELOPMENT OF STANDARDS

The Status of Training Prospective Teachers in Student Assessment

Schafer and Lissitz (1987) conducted a survey of AACTE member institutions in an attempt to determine their requirements for education students in the area of measurement. Responses were received from 438 of 707 institutions. The authors reported that “with the exception of school counseling and special education programs, 49% or more of the programs surveyed do not require for certification a formal course in measurement” (p. 61). Many of the institutions suggested measurement is covered in other courses that are required in their programs, but the authors questioned the value of measurement being taught incidentally and/or by professors who lack specific expertise in measurement.

Roeder (1972) conducted a survey of 940 elementary school teacher training institutions nationwide. Based on 860 usable responses, the author made the following observations:

While only 270 institutions reported requiring prospective elementary classroom teachers to complete a course devoted exclusively to tests and measures, 470 institutions required a course in play activities and games. 633 institutions reported requiring courses in music methods for classroom teachers, and 637 institutions required one or more courses in the art methods for classroom teachers. (p. 240)

Gullickson (1985) noted that colleges often provide some instruction in measurement and evaluation, but the time devoted to such instruction is limited. He observed, “Each professor is likely to choose topics he or she perceives as most important to teachers. As such, the professor’s choices will depend upon his or her knowledge of measurement” (p. 96).

In reviewing the literature on teacher knowledge of measurement, Farr and Griffin (1973) reached the following conclusions:

1. There should be concern over the adequacy of teacher preparation in administering, scoring, and interpreting standardized tests for that part of the vital role that teachers seem to play in testing. Also, though teachers have only minimal coursework in measurement, what should be the content of a tests and measurements course is a vital question that pre-service and in-service educators must face.

2. Teachers do not know much about measurement concepts particularly in relation to normative data and standardized tests. What they should know in terms of measurement concepts is another critical question.
3. Most studies of teachers’ measurement knowledge relate only to standardized tests and not classroom testing for planning instruction. Yet, teachers are using what knowledge they have of standardized tests to make critical decisions in regard to students’ academic programs.

4. Teachers occupy a central role in the testing and evaluation process of their pupils. They are deeply involved in testing, standardized or otherwise.

5. Standardized achievement test scores about pupils are relied on heavily by teachers and could have important effects on teachers’ attitudes and behaviors toward students, and might influence evaluations of classroom performance (e.g., Rosenthal Study [1968]). Teachers seem to have great faith in tests. (p. 23)

Rudman et al. (1980) published an extensive review of the literature on teacher preparation in assessment. In it they reported:

While there appears to be general agreement that teachers are not overly confident of their ability to interpret standardized test scores, the degree of confidence reported varies from researcher to researcher. Olejnik, (1979) in a study conducted among non-test specialists (counselors, teachers and building principals), found that over 90% of elementary and middle school educators indicated that they were at least “somewhat” confident of their ability to interpret test scores. The least confident were high school educationists. But when a mini-test similar to one given in college-level measurement courses was administered to the respondents, this self-reported “confidence” was not borne out. Most educationists correctly answered an item dealing with a percentile score (73%), yet a similar proportion missed an item that related norms to standards (77% incorrectly assumed that they were the same). They showed little understanding of the significance of stanine differences (only 35% recognized that a two stanine difference is significant), and very few could properly interpret a grade equivalent score (12%). On the basis of his study, Olejnik concluded that in spite of self-reported confidence it appeared that non-measurement specialists needed additional assistance in the interpretation of standard scores.

Stetz has conducted a series of studies aimed at determining the extent to which teachers and other educationists understand and accept standardized test results. His first study was a market survey of Stanford Achievement Test users (Stetz, 1977). Among a number of questions asked was one dealing with the types of scores they found most useful for assessment purposes. Both teachers and administrators reported that they preferred grade equivalents and percentile ranks for meeting their assessment needs; 59% of the teachers surveyed chose these two scores for individual student evaluation, 56% chose these two scores for class evaluation purposes, 65%
preferred these two scores for reporting test results to parents. One would like to assume from this that those who showed such a strong preference for these two standard scores understood what they signified, but Olejnik's study does give one some pause (Olejnik, 1979). (pp. 14-15)

Gullickson (1986) surveyed classroom teachers and professors responsible for teacher training to determine the measurement concepts viewed as important by the two groups. Gullickson reported strong disagreement between teachers and professors regarding statistics, nontest evaluation activities, and formative and summative evaluation:

Regarding statistics, two factors appear to be probable reasons for the teacher/professor disagreement. First, others who have assessed teachers' competency in measurement (see Rudman et al., 1980) have indicated that teachers do not have a good grasp of statistical concepts. This suggests that preservice measurement instruction, despite its relatively substantial emphasis on statistics, does not result in a level of understanding that would enable teachers to comfortably apply statistics to their evaluation needs. Such discomfort with statistics may well lead to devaluing of it.

Second, teachers may perceive such analyses as requiring more work than is justified by the benefits, particularly since statistical analyses can be avoided without obvious effect. In this regard, it is noteworthy that although there is substantial agreement among measurement experts as to the importance of statistical analyses, there is a paucity of empirical evidence to establish the positive instructional effects of such analyses.

Regarding nontest evaluation techniques, not only do professors give the topic substantially less emphasis than teachers recommend, but other research (Gullickson, 1985; Salmon-Cox, 1982; Stiggins & Bridgeford, 1982) indicates that teachers make substantial use of nontest evaluation techniques. Given their substantial use, greater emphasis on nontest evaluation techniques in preservice training programs should be expected. Here again there may be several reasons for the difference in professor and teacher opinions: (a) professors may not be aware of the extent to which teachers employ such techniques (research by Beck & Stetz, cited in Rudman et al., 1980, suggests that measurement experts do not have a clear understanding of teacher evaluation practices); (b) professors may perceive such techniques to be properly the domain of instructional methods courses and not the domain of measurement courses; and (c) professors may perceive the use of such techniques as less reliable and less valid than other evaluation techniques—thus deserving less emphasis.
Teacher and professor differences regarding formative and summative evaluation appear to stem from two possible roots. First, teacher priority items suggest that teachers recommend emphasis both on the general topics and on their specific applications. In contrast, professors give priority solely to the general issues with the expectation that specific applications will be provided in other methods courses. Certainly, given the diverse group of students who typically take an educational measurement course, presentation of examples appropriate to the needs of all students would be a time consuming and difficult task.

Second, five of the seven teacher-priority items for formative and summative evaluation relate directly to the identification and study of exceptional children (e.g., data to guide remediation, identifying gifted and slow learners, and identifying underachievers). None, however, was included among the professor priorities. This suggests that teachers alone place a high priority on the evaluation of special students. (pp. 350-353)

Perhaps Fleming (1979) addressed the teacher/professor conflicts most directly:

It appears that preservice teacher training with its emphasis on technical considerations and measurement processes as isolated events contribute to the ongoing dilemma for teachers. Is it too much to expect that training programs should foster a view of the instructional process as a continuum such as has been delineated by Tyler, for example, which in such a conceptualization consists of objectives, learning experiences and evaluation? (p. 2)

STANDARDS FOR TEACHER COMPETENCE IN EDUCATIONAL ASSESSMENT OF STUDENTS

By establishing standards for teacher competence in student assessment, the four involved associations subscribe to the view that student assessment is an essential part of teaching and that good teaching cannot exist without good student assessment. Training to develop the competencies covered in the standards should be an integral part of preservice preparation. Further, such assessment training should be widely available to practicing teachers through staff development programs at the district and building levels.

---

The committee that developed the standards represented four professional associations. James R. Sanders (Western Michigan University) chaired the committee and represented NCME along with John R. Hills (Florida State University) and Anthony J. Nitko (University of Pittsburgh). Jack C. Merwin (University of Minnesota) represented the American Association of Colleges for Teacher Education. Carolyn Trice represented the American Federation of Teachers. Marcella Dianda and Jeffrey Schneider represented the National Education Association. This section of the chapter represents the work of this committee and is a reproduction of the resulting document.
3. DEVELOPMENT OF STANDARDS

The standards are intended for use as:

- a guide for teacher educators as they design and approve programs for teacher preparation
- a self-assessment guide for teachers in identifying their needs for professional development in student assessment
- a guide for workshop instructors as they design professional development experiences for in-service teachers
- an impetus for educational measurement specialists and teacher trainers to conceptualize student assessment and teacher training in student assessment more broadly than has been the case in the past

The Approach Used to Develop the Standards

The memberships of the four associations are professional educators involved in teaching, teacher education, and student assessment. Members of these associations are concerned about the inadequate preparation of teachers for assessing the educational progress of their students, and thus sought to address this concern effectively. The committee named by the associations first met in September 1987 and affirmed its commitment to defining standards for teacher preparation in student assessment. The committee then undertook a review of the research literature to identify needs in student assessment, current levels of teacher training in student assessment, areas of teacher activities requiring competence in using assessments, and current levels of teacher competence in student assessment.

The members of the committee used their collective experience and expertise to formulate and then revise statements of important assessment competencies. Several drafts of these competencies were revised by the committee before the standards were released for public review. Comments by reviewers from each of the associations were then used to prepare this final statement.

Overview of the Standards

There were seven standards developed to cover assessment competencies needed by classroom teachers. In recognizing the critical need to revitalize classroom assessment, some standards focus on classroom-based competencies. Because of teachers' growing roles in education and policy decisions beyond the classroom, other standards address assessment competencies underlying teacher participation in decisions related to assessment at the school, district, state, and national levels.
The scope of a teacher’s professional role and responsibilities for student assessment may be described in terms of the following activities. These activities imply that teachers need competence in student assessment and sufficient time and resources to complete them in a professional manner:

- **Activities occurring prior to instruction:** (a) understanding students’ cultural backgrounds, interests, skills, and abilities as they apply across a range of learning domains and/or subject areas; (b) understanding students’ motivations and their interests in specific class content; (c) clarifying and articulating the performance outcomes expected of pupils; and (d) planning instruction for individuals or groups of students.

- **Activities occurring during instruction:** (a) monitoring pupil progress toward instructional goals; (b) identifying gains and difficulties pupils are experiencing in learning and performing; (c) adjusting instruction; (d) giving contingent, specific, and credible praise and feedback; (e) motivating students to learn; and (f) judging the extent of pupil attainment of instructional outcomes.

- **Activities occurring after the appropriate instructional segment (e.g., lesson, class, semester, grade):** (a) describing the extent to which each pupil has attained both short- and long-term instructional goals; (b) communicating strengths and weaknesses based on assessment results to students and parents or guardians; (c) recording and reporting assessment results for school-level analysis, evaluation, and decision making; (d) analyzing assessment information gathered before and during instruction to understand each student’s progress to date and to inform future instructional planning; (e) evaluating the effectiveness of instruction; and (f) evaluating the effectiveness of the curriculum and materials in use.

- **Activities associated with a teacher’s involvement in school building and school district decision-making:** (a) serving on a school or district committee examining the school’s and district’s strengths and weaknesses in the development of its students; (b) working on the development or selection of assessment methods for school building or school district use; (c) evaluating school district curriculum; and (d) other related activities.
• Activities associated with a teacher's involvement in a wider community of educators: (a) serving on a state committee asked to develop learning goals and associated assessment methods; (b) participating in reviews of the appropriateness of district, state, or national student goals and associated assessment methods; and (c) interpreting the results of state and national student assessment programs.

Each standard that follows is an expectation for assessment knowledge or skill that a teacher should possess in order to perform well in the five areas just described. As a set, the standards call on teachers to demonstrate skill in selecting, developing, applying, using, communicating, and evaluating student assessment information and student assessment practices. A brief rationale and illustrative behaviors follow each standard.

The standards represent a conceptual framework or scaffolding from which specific skills can be derived. Work to make these standards operational will be needed even after they have been published. It is also expected that experience in the application of these standards should lead to their improvement and further development.

The Standards

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.

Skills in choosing appropriate, useful, administratively convenient, technically adequate, and fair assessment methods are prerequisite to good use of information to support instructional decisions. Teachers need to be well acquainted with the kinds of information provided by a broad range of assessment alternatives and their strengths and weaknesses. In particular, they should be familiar with criteria for evaluating and selecting assessment methods in light of instructional plans.

Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to use the concepts of assessment error and validity when developing or selecting their approaches to classroom assessment of students. They will understand how valid assessment data can support instructional activities such as providing appropriate feedback to students, diagnosing group and individual learning needs, planning for individualized educational programs, motivating students, and evaluating instructional procedures. They will understand how invalid information can affect
instructional decisions about students. They will also be able to use and evaluate assessment options available to them, considering among other things, the cultural, social, economic, and language backgrounds of students. They will be aware that different assessment approaches can be incompatible with certain instructional goals and may impact quite differently on their teaching.

Teachers will know, for each assessment approach they use, its appropriateness for making decisions about their pupils. Moreover, teachers will know where to find information about and/or reviews of various assessment methods. Assessment options are diverse and include text- and curriculum-embedded questions and tests, standardized criterion-referenced and norm-referenced tests, oral questioning, spontaneous and structured performance assessments, portfolios, exhibitions, demonstrations, rating scales, writing samples, paper-and-pencil tests, seatwork and homework, peer- and self-assessments, student records, observations, questionnaires, interviews, projects, products, and others’ opinions.

2. **Teachers should be skilled in developing assessment methods appropriate for instructional decisions.**

While teachers often use published or other external assessment tools, the bulk of the assessment information they use for decision making comes from approaches they create and implement. Indeed, the assessment demands of the classroom go well beyond readily available instruments.

Teachers who meet this standard will have the conceptual and application skills that follow. Teachers will be skilled in planning the collection of information that facilitates the decisions they will make. They will know and follow appropriate principles for developing and using assessment methods in their teaching, avoiding common pitfalls in student assessment. Such techniques may include several of the options listed at the end of the first standard. The teacher will select the techniques which are appropriate to the intent of the teacher's instruction.

Teachers meeting this standard will also be skilled in using student data to analyze the quality of each assessment technique they use. Since most teachers do not have access to assessment specialists, they must be prepared to do these analyses themselves.

3. **Teachers should be skilled in administering, scoring, and interpreting the results of both externally-produced and teacher-produced assessment methods.**

It is not enough that teachers are able to select and develop good assessment methods; they must also be able to apply them properly.
Teachers should be skilled in administering, scoring, and interpreting results from diverse assessment methods.

Teachers who meet this standard will have the conceptual and application skills that follow. They will be skilled in interpreting informal and formal teacher-produced assessment results, including pupils’ performances in class and on homework assignments. Teachers will be able to use guides for scoring essay questions and projects, stencils for scoring response-choice questions, and scales for rating performance assessments. They will be able to use these in ways that produce consistent results.

 Teachers will be able to administer standardized achievement tests and be able to interpret the commonly reported scores: percentile ranks, percentile band scores, standard scores, and grade equivalents. They will have a conceptual understanding of the summary indexes commonly reported with assessment results: measures of central tendency, dispersion, relationships, reliability, and errors of measurement.

Teachers will be able to apply these concepts of score and summary indices in ways that enhance their use of the assessments that they develop. They will be able to analyze assessment results to identify pupils’ strengths and errors. If they get inconsistent results, they will seek other explanations for the discrepancy or other data to attempt to resolve the uncertainty before arriving at a decision. They will be able to use assessment methods in ways that encourage students’ educational development and that do not inappropriately increase students’ anxiety levels.

4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.

Assessment results are used to make educational decisions at several levels: in the classroom about students, in the community about a school and a school district, and in society, generally, about the purposes and outcomes of the educational enterprise. Teachers play a vital role when participating in decision making at each of these levels and must be able to use assessment results effectively.

Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to use accumulated assessment information to organize a sound instructional plan for facilitating students’ educational development. When using assessment results to plan and/or evaluate instruction and curriculum, teachers will interpret the results correctly and avoid common misinterpretations, such as basing decisions on scores that lack curriculum
validity. They will be informed about the results of local, regional, state, and national assessments and about their appropriate use for pupil, classroom, school, district, state, and national educational improvement.

5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.

Grading students is an important part of professional practice for teachers. Grading is defined as indicating both a student’s level of performance and a teacher’s valuing of that performance. The principles for using assessments to obtain valid grades are known and teachers should employ them.

Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to devise, implement, and explain a procedure for developing grades composed of marks from various assignments, projects, in-class activities, quizzes, tests, and/or other assessments that they may use. Teachers will understand and be able to articulate why the grades are rational, justified, and fair, acknowledging that such grades reflect their preferences and judgments. Teachers will be able to recognize and to avoid faulty grading procedures such as using grades as punishment. They will be able to evaluate and to modify their grading procedures in order to improve the validity of the interpretations made from them about students’ attainments.

6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.

Teachers must routinely report assessment results to students and to parents or guardians. In addition, they are frequently asked to report or to discuss assessment results with other educators and with diverse lay audiences. If the results are not communicated effectively, they may be misused or not used. To communicate effectively with others on matters of student assessment, teachers must be able to use assessment terminology appropriately and must be able to articulate the meaning, limitations, and implications of assessment results. Furthermore, teachers will sometimes be in a position that will require them to defend their own assessment procedures and their interpretations of them. At other times, teachers may need to help the public to interpret assessment results appropriately.

Teachers who meet this standard will have the conceptual and application skills that follow. Teachers will understand and be able to give appropriate explanations of how the interpretation of student assessments must be moderated by the student’s socioeconomic,
cultural, language, and other background factors. Teachers will be able to explain that assessment results do not imply that such background factors limit a student’s ultimate educational development. They will be able to communicate to students and to their parents or guardians how they may assess the student’s educational progress. Teachers will understand and be able to explain the importance of taking measurement errors into account when using assessments to make decisions about individual students. Teachers will be able to explain the limitations of different informal and formal assessment methods. They will be able to explain printed reports of the results of pupil assessments at the classroom, school district, state, and national levels.

7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

Fairness, the rights of all concerned, and professional ethical behavior must undergird all student assessment activities, from the initial planning for and gathering of information to the interpretation, use, and communication of the results. Teachers must be well versed in their own ethical and legal responsibilities in assessment. In addition, they should also attempt to have the inappropriate assessment practices of others discontinued whenever they are encountered. Teachers should also participate with the wider educational community in defining the limits of appropriate professional behavior in assessment.

Teachers who meet this standard will have the conceptual and application skills that follow. They will know those laws and case decisions that affect their classroom, school district, and state assessment practices. Teachers will be aware that various assessment procedures can be misused or overused, resulting in harmful consequences such as embarrassing students, violating a student’s right to confidentiality, and inappropriately using students’ standardized achievement test scores to measure teaching effectiveness.

WHERE DO WE GO FROM HERE?

In 1986, after studying the feasibility of the NCME taking on a licensing or certifying (i.e., credentialing) role for measurement experts, it was noted that the nature of measurement expertise in education was too illusory ever to be able to define, or standardize, requirements across the education profession. Instead, collaborative studies with professional education associations were planned to
identify the assessment competencies needed to perform in different professional roles, and to prepare joint statements about the preservice and in-service preparation in student assessment of educators filling these different roles.

The classroom teacher role was the first to be studied. The resulting standards are intended to be a statement that will affect teacher certification requirements and the accreditation of teacher preparation programs. There is an expectation that administrator, counselor, testing director, special education director, curriculum director, and other roles will require similar attention in the future.

Now that the teacher standards have been developed, there are a number of follow-up activities that deserve the attention of the four collaborating associations. These include:

- collaborating on a table of specifications for each standard, and then developing assessment procedures and instruments for assessing the extent to which an individual can meet the standards.
- collaborating on instructional modules and workshops for teachers based on the standards.
- collaborating on developing a curriculum strand to prepare preservice teachers for student assessment. This curriculum strand might contain grounded scenarios of classroom teaching in which teachers are meeting and not meeting the standards, with analyses and instruction to accompany each scenario.
- collaborating on the dissemination and use of the standards through the four associations, state departments of education, and such projects as the National Board of Teaching.

Another thrust for the future would be for the NCME to work with the American Association of School Administrators (AASA), the National Association of Secondary School Principals (NASSP), and the National Association of Elementary School Principals (NAESP) to prepare similar standards for school administrators. This pattern of collaborative development could then continue for educator groups that include testing directors, counselors, special education specialists, curriculum specialists, and other professional groups that might be added. By the time standards and spinoff products are developed and are being used for each of these groups, it would then be time to review and update each set of standards in a collaborative and systematic manner. A review by the cooperating associations every 5 years would be in order.
There is still a great deal of work to be done to improve the quality of student assessments in education. The first step taken by the four associations to develop these standards for teacher competence in student assessment is a major step in the right direction.

REFERENCES


INTRODUCTION

In his role as a discussant of a series of papers on educational evaluation 23 years ago, Scriven (1970) made the following comments:

While the papers this afternoon did not, on the above account, go far enough in the direction of basic evaluation, from another point of view they began at too abstract a level. They contain no discussion at all of the basic method of educational evaluation, one whose use quantitatively swamps any other. I refer to the practice of grading. Like so many other everyday practices, grading has often seemed too humble to merit the attention of high-powered test and measurement people. My feeling is that it is far more important and in more need of help than anything else they work on. Moreover it admirably illustrates the point just made, that the new critics of bad practices are about as irrational as most defenders of the practices.

Unfortunately, little has changed since this observation was made.

Reference Works on Educational Measurement and Research

A brief review of three standard reference works reveals a general disdain for the topic of grading. The recently published third edition of Educational Measurement (Linn, 1989) contains two chapters that
might logically be expected to touch on grading. Chapter 12, entitled “Designing Tests That Are Integrated with Instruction,” identifies attainment decisions as one of four types of decisions for which tests are employed. The author devotes approximately one-half page (out of 24 in the chapter) to this type of decision and never mentions grading in relation to attainment. Chapter 14, entitled “Certification of Student Competence,” provides a lengthy review of statewide competency testing programs and issues associated with standard setting in such programs. The author has nothing to say about the teacher’s role in the certification of competence and standard setting as it relates to grading.

Apparently it simply doesn’t occur to measurement specialists that classroom teachers are the ones who have the primary responsibility for making attainment decisions and certifying student competence. The terms grades and grading do not appear in the index of *Educational Measurement* (Linn, 1989).

A second standard reference is the third edition of the *Handbook of Research on Teaching* (Wittrock, 1986). The three chapters in this volume that would logically be linked to grading practices are Chapter 13 (“Teaching Functions”), Chapter 14 (“Classroom Organization and Management”), and Chapter 17 (“Philosophy of Teaching”). None of these chapters contains any reference to grades.

A third somewhat more general reference is the most recent *Encyclopedia of Educational Research* (Mitzel, 1982). In this volume there are approximately 10 pages devoted to the topic Marking Systems. As the title suggests, this summary deals primarily with the purposes of marking and the popularity of various marking systems. The only reference to the process of assigning grades is one page that addresses various orientations (criterion referenced, norm referenced, student potential) a teacher may adopt in determining grades. The orientation a teacher adopts is clearly a topic with both philosophic and psychometric importance. (More will be said about this later.) However, the review in the *Encyclopedia* deals primarily with the relative popularity of these orientations as revealed in surveys of teachers.

**Textbooks on Classroom Measurement and Evaluation**

A second potential source of information on grading is the textbooks that provide the framework for the education of teachers on matters related to classroom evaluation. Because teachers are almost universally required to assign grades to students and because these grades are commonly defined to reflect the teacher’s evaluation of the performance of students on various tests, quizzes, etc. designed by the
teacher, it follows that textbooks on classroom assessment should provide a wealth of practical advice on how to assign grades to students. Alas, such is not the case!

A sample of 12 such texts was examined. This is not a random sample. Rather, it represents all such texts that were easily accessible. It is likely that this set is biased in favor of texts that are most commonly adopted, due to the fact that 5 of the texts have gone through at least three editions.

Table 1 presents a summary that identifies the texts and gives information concerning the length of each and the number of pages on grading. All texts except one (Hills, 1981) contain a single chapter on a variety of issues associated with grading and grading systems. The number of pages in this chapter in relation to the total length of the book is typically quite small, ranging from 4% to 10%. (For the six chapters in Hills, 1981, the figure is 22%.) As shown in the last column of the table, the number of pages devoted to the actual process of assigning grades (as opposed to discussions of various grading and reporting systems) is pitifully small. Only two authors (Hills and Carey) devote more than 10 pages to the actual grading process and half the books devote only 5 or 6 pages to the topic. It seems fair to conclude that, with two possible exceptions, authors of these textbooks on classroom measurement do not attach a great deal of importance to providing teachers with practical advice on grading.

Table 1. Summary of Treatment of Grading in "Standard" Texts on Educational Measurement

<table>
<thead>
<tr>
<th></th>
<th>Edition/Year</th>
<th>Total Pages (Excluding Appendices)</th>
<th>Pages in Grading Chapter</th>
<th>Pages Devoted to Grade Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>a Ahman &amp; Glock</td>
<td>5th/1975</td>
<td>430</td>
<td>40</td>
<td>6</td>
</tr>
<tr>
<td>Carey</td>
<td>1st/1988</td>
<td>415</td>
<td>40</td>
<td>18</td>
</tr>
<tr>
<td>Ebel &amp; Frisbie</td>
<td>4th/1986</td>
<td>340</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>Gronlund</td>
<td>5th/1985</td>
<td>488</td>
<td>26</td>
<td>9</td>
</tr>
<tr>
<td>b Hills</td>
<td>2nd/1981</td>
<td>380</td>
<td>84</td>
<td>25</td>
</tr>
<tr>
<td>Hopkins &amp; Antes</td>
<td>2nd/1985</td>
<td>465</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>Hopkins, Stanley, &amp; Hopkins</td>
<td>7th/1990</td>
<td>470</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Kubisyn &amp; Borich</td>
<td>2nd/1987</td>
<td>430</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>Mehrens &amp; Lehmann</td>
<td>3rd/1984</td>
<td>595</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>Nitko</td>
<td>1st/1983</td>
<td>585</td>
<td>24</td>
<td>6</td>
</tr>
<tr>
<td>a,c Noll, Scannell, &amp; Craig</td>
<td>4th/1979</td>
<td>480</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>c Popham</td>
<td>2nd/1990</td>
<td>395</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

\( ^a \) The most recent edition of this text was not available for review.

\( ^b \) Hills devotes six chapters to various issues associated with grades and grading.

Three chapters deal with the actual grading process.

\( ^c \) Grading is covered in a general chapter on the uses of data.
Research Literature on Measurement

A possible final source of advice on grading is the general literature on educational measurement and/or research. An ERIC search was performed covering the literature for the period from January 1, 1976 through September 30, 1989. A total of 91 references was obtained using the descriptor “Assigning Grades.” A careful reading of the abstracts for these 91 references revealed that over half of them (54) did not address or dealt only marginally with assigning grades to students in classroom settings. For example, many of these focus upon issues of evaluating student performance in specific settings (rating systems for college-level writing assignments, using reading journals to improve comprehension of complex texts, etc.) or general student evaluation issues (policies on homework assignments in secondary schools, testing practices of teachers in specific educational settings, etc.). Others deal primarily with curriculum issues, the relationship of grades to student ratings of teachers, etc.

The 37 remaining articles can be classified according to the type of article (empirical study, critique/recommendation) and the educational level (Grades K-12, Postsecondary, Unspecified) to which it is addressed. The results are shown in Table 2. There are two striking features revealed in this table. First, the empirical studies of grading are outnumbered by articles that either critique or recommend grading practices by a 2:1 ratio. Second, half the articles refer to grading at the postsecondary level, and the remaining half are equally split between those that refer to precollege settings and articles that are general with respect to educational level.

The numbers in parentheses in the first column of Table 2 refer to the number of survey studies. These studies typically report results based upon responses of teachers in a small group of educational institutions. In each case they employ a self-report instrument designed to determine the popularity of various grading philosophies and practices. Survey results at both the secondary (Terwilliger, 1987) and college level (Prather, Smith, & Kodras, 1981) consistently reveal differences in grading philosophies and practices as a function of the subject matter field.

The differences among disciplines are even more obvious when one examines the articles that focus upon critiques and recommendations related to grading. Seven of the 12 articles at the postsecondary

---

Several other descriptors were employed before selecting this phrase. These resulted in extensive lists of references, most of which have nothing to do with the topic of grading (e.g., using the descriptor “Grades” results in 8,547 references, mostly dealing with research on different grade levels in public schools).
level and two of the four articles at the K-12 level address grading practices within specific subject matter fields. In each case the author critiques practices or recommends alternative grading strategies that are somewhat unique to instructional methods employed in that field. These range from articles on grading in algebra and engineering courses to courses on personal development and career planning. Two articles (Calhoun & Beattie, 1984; Cohen, 1983) deal specifically with grading practices appropriate for special education students who are in mainstream classes. Advice on how to assign grades in such special circumstances currently is not found in standard texts on classroom measurement and evaluation. The nine articles that are not specific with respect to educational level tend to focus either upon narrow technical issues such as determining boundaries for grading (Aiken, 1983), using computers in assigning grades (Hsiao, 1985), or innovative approaches to grading such as contracts (Klein, 1976).

It would be futile to attempt to synthesize the findings and recommendations offered in the 37 articles in Table 2. The literature on grading is defined more by its diversity than by any universal themes. Differences between educational levels and subject matter fields make generalizations risky, if not meaningless. Yet one gets the sense that the fundamental issues at the heart of grading practices are philosophic, not psychometric, in nature. Perhaps this is why the "high-powered test and measurement people" that Scriven (1970) referred to have so little to say on the subject. Therefore, it may be wise to turn elsewhere for perspectives that can, and often do, influence teachers' grading practices.

Table 2. Summary of Articles on Assigning Grades

<table>
<thead>
<tr>
<th>Educational Level</th>
<th>Type of Article</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Empirical Study</td>
</tr>
<tr>
<td>Grade Levels K-12</td>
<td>5 (1)</td>
</tr>
<tr>
<td>Post Secondary</td>
<td>7 (4)</td>
</tr>
<tr>
<td>General/ Unspecified Level</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>12 (5)</td>
</tr>
</tbody>
</table>

It would be futile to attempt to synthesize the findings and recommendations offered in the 37 articles in Table 2. The literature on grading is defined more by its diversity than by any universal themes. Differences between educational levels and subject matter fields make generalizations risky, if not meaningless. Yet one gets the sense that the fundamental issues at the heart of grading practices are philosophic, not psychometric, in nature. Perhaps this is why the "high-powered test and measurement people" that Scriven (1970) referred to have so little to say on the subject. Therefore, it may be wise to turn elsewhere for perspectives that can, and often do, influence teachers' grading practices.
TRADITIONAL GRADING\(^2\) AND PHILOSOPHIC ORIENTATIONS

It is not possible to discuss traditional grading practices in an informed manner without first examining the set of beliefs and assumptions underlying such practices. This is rarely done by advocates of traditional approaches to grading (e.g., authors of textbooks on classroom evaluation). However, philosophic views are discussed at length by a variety of critics of grading, both within and outside the professional educational establishment. Because the views of these critics are not without merit and have a great intuitive appeal to many teachers, they should be examined carefully. Consider the following questions:

1. What purposes do grades serve?
2. What are the costs and benefits of grades?
   a. To students
   b. To society
3. On what basis should students be judged?
   a. What data are relevant?
   b. How should the data be evaluated?

Advocates of Traditional Grading

To the question concerning the purposes served by grades, those who support them would likely give two answers. First, grades provide a useful basis for making a variety of important decisions by (and about) individual students. These might include (a) determining promotion and/or graduation, (b) awarding scholarships or special honors, (c) determining eligibility for special programs for the talented, and (d) determining admission into college or other advanced training. Second, grades provide a tangible recognition for excellence in academic pursuits. Such recognition rewards past efforts and encourages future success in learning.

Gardner (1984) has described U.S. education as a sorting-out process:

Americans believe that promise should be recognized at whatever level in society it occurs. They like to think that those future presidents dashing off to school may come from any walk of life.

But as education becomes increasingly effective in pulling the able youngster to the top, it becomes an increasingly rugged sorting-out

\(^2\)Grading is defined here as the process by which a teacher arrives at a value judgment concerning the quality of a student's achievement of course objectives during a specified period of instruction. Evaluation of performances on single examinations, assignments, projects, etc., are discussed in other papers in this volume.
4. SOME THOUGHTS ON GRADING

process for everyone concerned. The schools are the golden avenue of opportunity for able youngsters; but they are also the arena in which less able youngsters discover their limitations. This thought rarely occurred to the generations of Americans who dreamed of universal education. They saw the beauty of a system in which young people could go as far as their ability and ambition would take them, without obstacles of money, social standing, religion, or race. They didn’t reflect on the pain involved for those who lacked the necessary ability. Yet pain there is and must be. (p. 79)

With regard to costs and benefits, advocates of traditional grades state that the sorting process which results from grading, although sometimes painful, is ultimately of benefit both to students and to society as a whole. Although grades are admittedly imperfect, they do provide an important basis for a meritocracy. Moynihan (1971) stated this succinctly when he commented:

One of the achievements of democracy, although it seems not much regarded as such today, is the system of grading and sorting individuals so that young persons of talent born to modest or lowly circumstances can be recognized for their worth. (Similarly, it provides a means for young persons of social status to demonstrate that they have inherited brains as well as money, as it were.) I have not the least doubt that this system is crude, that it is often cruel, and that it measures only a limited number of things. Yet it measures valid things, by and large. To do away with such systems of accreditation may seem like an egalitarian act, but in fact it would be just the opposite. We would be back to a world in which social connections and privilege count for much more than any of us, I believe, would like. If what you know doesn’t count, in the competitions of life, who you know will determine the outcomes. (p. 4)

It is generally agreed among advocates of grading (at least those who write textbooks on measurement) that the basis for a grade should be the performance or achievement of a student, not the effort expended, work habits, character traits, etc. The reason for keeping the basis for grades as “pure” as possible is to minimize the confusion that arises when the meaning of a grade is interpreted. A separate system for recording and reporting teacher judgments concerning student effort, work habits, character traits, etc. is recommended if a school system decides such information is desirable.3

3There is a practical question of how many judgments a teacher should be expected to make and how reliable such judgments are likely to be. This may differ substantially depending upon the setting (e.g., primary grade self-contained classes vs. secondary school classes).
It is further agreed by most advocates of grading that grades should reflect a judgment of achievement with respect to other students (i.e., grades should be norm referenced). This is consistent with the belief that a primary purpose of grades is to differentiate among students as part of an ongoing sorting and decision-making process. This is nicely summarized in the following quote from a colleague who served on a student/faculty committee charged with examining the grading system at the University of Minnesota:

In education, grading represents an information system. Historians perhaps can tell us whether the idea of grades originated from the needs of teachers or the needs of pupils. Current critics can comment on the pro-grading motivation of some administrators and the anti-grading motivation of some students. Such commentary, historical or contemporary, seems not to contribute much to logical analysis. The present social climate encourages a view of academic grading as pejoratively “discriminatory” rather than helpfully “discriminating.” The ultimate reality is that Nature does differentiate. Given that fact, we may retreat philosophically from the ensuing pejorative “competition,” or we can advance functionally with a helpful “division of labor.” (Schofield, 1972.)

Finally, with regard to alternatives (e.g., narrative reports, parent-teacher conferences, contract grading, etc.), advocates of traditional grades consider these to be generally impractical due to time demands that they place on both teachers and those who typically employ grades in decision making. It should be noted, however, that the feasibility of alternatives to traditional grading depends upon the educational context. This will be discussed at greater length in a later section of this paper.

Critics of Traditional Grading

There are many critics of traditional grading. Three identifiable groups will be discussed. The first comprises individuals who identify strongly with the humanistic movement in education. During the 1960s and the 1970s they advocated fundamental changes in the structure of education and the organization of schools. This movement gave birth to a variety of open or alternative schools in many parts of the United States. A series of publications by Kirschenbaum, Simon, and Napier (1971), Simon and Bellanca (1976), and Bellanca (1977) deal specifically with problems associated with traditional grading and describe alternatives that are thought to be superior to it.

A second source of criticism of traditional grading practices comes from social psychologists and educators who have analyzed
educational practices from the perspective of cooperation versus competition. Deutsch (1979), Johnson and Johnson (1974), and Slavin (1977) have argued that classroom evaluation and reward structures that foster competition among students create an unhealthy environment for learning. They advocate classroom organizations based upon student groups that emphasize teamwork and cooperative learning strategies.

The philosophical premises of this perspective are variants upon humanistic themes. Perhaps the clearest critique of the traditional view of society and grading has been offered by Deutsch (1979):

In addition, I believe we must begin to challenge the assumptions underlying the competitive, meritocratic ideology of our society. We must question whether socioeconomic position in our society is actually distributed on the basis of individual merit. In addition, we must raise issue with the notion that merit belongs solely to an individual, as though its possession were not strongly influenced by social and biological circumstances largely beyond the individual’s control. And we must raise doubts about the traditional answer to the question, Who merits merit?—namely, those who have most merit as a consequence of having been more favored with the conditions that foster merit. Finally, we must raise the central question: If the competitive grading system in our schools—a less corrupted version of a competitive merit system than the one that characterizes our larger society—does not foster a social environment that is conducive to individual well-being and effective social cooperation, why would one expect that such values would be fostered in a society that is dominated by a competitive, meritocratic ideology? If the competitive-hierarchical atmosphere is not good for our children, is it good for us? (p. 401)

Research reviews by Johnson and Johnson (1974) and Slavin (1977) conclude that cooperative learning strategies produce achievement outcomes equal to or better than competitive learning approaches in many classroom settings. Further, they conclude that student attitudes toward school and toward peers is much more positive in cooperative learning environments. It should be noted that most of these studies were conducted in elementary schools.

A third group of critics of traditional educational practices has become active in the outcome-based school movement. As reflected in a statement by Spady (1981), this group adopts a strong behavioristic approach to education with an emphasis upon detailed and explicit statements of learning outcomes, mastery-based instructional systems, and criterion-referenced assessment procedures. Spady (1981) lists the following philosophical premises of outcome-based education:
1. Almost all students are capable of achieving excellence in learning the essentials of formal schooling.
2. Success influences self-concept; self-concept influences learning and behavior.
3. The instructional process can be changed to improve learning.
4. Schools can maximize the learning conditions for all students by:
   a. establishing a school climate which continually affirms the worth and diversity of all students;
   b. specifying expected learning outcomes;
   c. expecting that all students perform at high levels of learning;
   d. ensuring that all students experience opportunities for personal success;
   e. varying the time for learning according to the needs of each student and the complexity of the task;
   f. having staff and students both take responsibility for successful learning outcomes;
   g. determining instructional assignment directly through continuous assessment of student learning; and
   h. certifying educational progress whenever demonstrated mastery is assessed and validated. (p. 2)

As might be expected, none of the three groups of critics believe that traditional grades serve a useful purpose. Grades are viewed as an artificial and harmful reward system that has little to do with learning. Grades are also seen as a mechanism to exert control over students. Students who learn to please the teacher are rewarded with high grades; students who do not frequently suffer low self-esteem and quit trying. Furthermore, even if grades reflect general learning, they provide little or no information concerning specifically what a student has learned.

Critics argue that the costs of traditional grading both to students and to society as a whole far outweigh the benefits. They claim that the disruptive effect of grades upon the educational process cannot be justified by the rather weak relationship of grades to later educational success, although it is admitted that secondary school grades are the best single predictor of college grades. The strength of the typical correlation between secondary and college grades (e.g., .50-.60) is not regarded as having any practical utility. The lack of any systematic relationship between grades and indices of success in nonschool settings (i.e., on-the-job performance) is also frequently noted.
Critics vary somewhat with respect to their preference for the proper basis for evaluating students. Humanists are proponents of approaches to evaluation that incorporate as much information as possible about the individual student. For example, they typically recommend that student achievement be judged with respect to the ability or improvement that a student demonstrates. Thorndike (1969a) has referred to this as evaluation with respect to potential. Another approach is "grading by contract." All these approaches individualize the judgment made by teachers and virtually assures all students who made a reasonable effort that they will not fail.

Advocates of cooperative learning strategies are highly critical of norm-referenced assessment and grading, which they regard as the epitome of a competitive system. As an alternative they emphasize group projects in which the assessment of each individual student is heavily dependent upon the quality of the product produced by the student's group. Other factors that determine a student's evaluation might include ratings by peers within the student's group, ratings by peers who are not members of the student's group, teacher observations of group interactions, and selected individual achievement data that are independent of group data. The relative weighting of each of these factors varies from one setting to the next. However, the important point is that the grade assigned to each student is influenced by both the performance of the team and the members' perceptions of the contributions made by the student to the team's success.

Advocates of outcome-based education also reject the norm-referenced sorting of students associated with traditional grading. Instead, they propose specific a priori statements of learning outcomes against which student performance can be judged. They argue that detailed publicly stated goals provide a more informative basis for evaluation. The criterion-referenced system associated with outcome-based education also is often linked with mastery learning approaches that provide students with multiple trials to demonstrate their competencies. General guidelines for establishing such a system are given by Spady (1981).

Some years ago Ebel (1974) listed 22 arguments (including those cited here) frequently made by the critics of traditional grades. He briefly analyzed each argument and presented a rebuttal. A summary of 8 of the most basic arguments and rejoinders given by Ebel is shown in Figure 1. Readers who wish to pursue this further are encouraged to read Ebel's article in its entirety.
<table>
<thead>
<tr>
<th>Criticism</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A single symbol cannot possibly report adequately the complex details of an educational achievement.</td>
<td>Grades aren't intended to provide details. They represent a method of reporting value judgments regarding general level of achievement.</td>
</tr>
<tr>
<td>2. The most important outcomes are intangible and hence cannot be assessed or graded.</td>
<td>Important outcomes are, by definition, those that make a difference. With properly constructed measuring devices, differences can be detected and can be the basis for grades.</td>
</tr>
<tr>
<td>3. Grades are ineffective motivators of real achievement in education.</td>
<td>Research studies indicate that differential grading does tend to motivate students. It is misleading to imply that high grades and &quot;real achievement&quot; are incompatible. When grades are properly given they are parallel.</td>
</tr>
<tr>
<td>4. When students learn mastery, as they should, no differential levels of achievement remain to be graded.</td>
<td>Mastery is difficult to define and does not insure identical levels and types of achievement. In almost any instructional setting some students learn faster and more than do others. This should be reflected in the grade reporting.</td>
</tr>
<tr>
<td>5. Low grades may discourage the less able pupils from efforts to learn. Also, some pupils will inevitably fail.</td>
<td>While there can be no guarantee that pupils will not receive low grades, special tutorial and remedial help should be offered to those who receive low grades. No pupil who has taken advantage of such help and made a serious effort to learn should be failed.</td>
</tr>
<tr>
<td>6. Grades set universal standards for all pupils despite their great individual differences.</td>
<td>A thoughtful teacher will set standards which are realistic for the class so that the highest grades are achievable. Individual differences in grades are intended to reflect important differences among students.</td>
</tr>
<tr>
<td>7. Grading fosters competition rather than cooperation.</td>
<td>Grading emphasizes individual achievement but that does not necessarily imply a competitive learning environment. Many students achieve individual excellence through cooperative learning activities.</td>
</tr>
<tr>
<td>8. Grading is more compatible with subject-centered education than with humanistic, child-centered education.</td>
<td>The distinction between subject-centered and childcentered education is not valid. A teacher can recognize his pupils as unique human beings and also help them to achieve subject matter objectives.</td>
</tr>
</tbody>
</table>
The Temporal Factor

Like all other educational practices, grading practices are influenced by fads and fashions. There are clear cyclical changes in such matters as the choice of the grading system (percent scale, letter grades, pass/fail, etc.) to employ. This is well documented on a general level by Cureton (1971) and in a specific setting by Wrinkle (1947). Little can be learned about the process of grading by studying the popularity of grading systems at any given point in time. The number of categories in grading systems and the symbols that are used may change with time, but these represent somewhat superficial concerns.

On a different level, the influence of various philosophical positions ebbs and flows with the passage of time. The alternative school movement associated with the humanistic view of education became very prominent during the late 1960s and 1970s. Consequently, there was much greater attention during that period to alternative grading practices advocated by humanistic educators. Many schools and colleges modified their grading systems (e.g., replacing "Failure" [F] with "No Credit" [N], providing "Satisfactory/No Credit" [S/N] as an option to letter grades, etc.) and the grade inflation phenomenon was born. For many students, grades were regarded as irrelevant.

More recently, the pendulum has swung back toward a more traditional view. Many of the modifications that were introduced as reforms 20 years ago have been replaced by systems that bear a striking similarity to those that were in place prior to 1960. S/N grading is now less popular and the F has been resurrected in many institutions. In response to grade inflation, a more refined grading system (A+, A, A-, etc.) has been adopted by some colleges in an effort to better differentiate among students. Grades now seem to be regarded as more important by students than they were 20 years ago.

Gardner (1984) has described the situation succinctly in discussing the continuing debate over demands for educational excellence vs. demands for educational equality. Although not identical to issues in grading controversies, there is a substantial overlap in philosophical viewpoints:

If the swings of the pendulum have been excessive at times and the debate more embittered than one might wish, it is because there are extreme and polarizing elements on both sides of the debate.
On the side of quality, the best proponents care deeply about standards and solid subject matter, seek to challenge and stretch the student, and believe that with appropriate adjustments these are suitable goals for students at every level of ability. Unfortunately, also on the side of quality are some who really care only about the college preparatory students and (whether they admit it or not) look down on all the others. Not surprisingly, they give an unpleasant tone to the debate.

On the side of equality, the best proponents care deeply about the economically deprived and about the student of lesser ability—but fully recognize the need for rigorous college preparatory programs. Unfortunately, also on the side of equality are some who are profoundly anti-intellectual, anti-subject matter, and anti-discipline. (p. 89)

The Educational Level Factor

A critical, but frequently ignored, variable in discussing grading systems and practices is the educational level of the students being evaluated. The importance of educational level follows from the fact that the number and types of decisions made by (and about) students change in significant ways, depending upon the educational and developmental stage at which a student is functioning. The impact of grading upon students is also likely to be different for students at different stages of maturity. For present purposes, four educational levels will be considered: grades K-6, grades 7-12, undergraduate college (13-16), and postgraduate level (e.g., graduate school, medical or law school, other advanced educational programs). Each of these four will be considered briefly.

At the earliest stages (grades K-6) in the educational process, the decisions that are made concerning a pupil’s educational progress are very limited. The primary question is, “Has this pupil acquired the basic knowledge and skills typically expected of children at this level?” If the answer is “yes,” the decision is to promote the individual to the next level. If the answer is “no,” a variety of actions are possible, depending upon the resources available (e.g., do not pro-

---

*Naturally, it is assumed that it is reasonable to expect the student to make typical progress. If there is evidence of a serious limitation upon the ability of a child to learn (e.g., certain physical or mental handicapping conditions), it is pointless to hold expectations of typical progress. Under these circumstances, the teacher needs to develop a separate set of expectations that are appropriate to the particular setting. The evaluation of students in such special educational circumstances relies heavily upon judgments of “progress with respect to potential.”*
mote the pupil and repeat the entire year of instruction, provide intensive remedial instruction during the summer as a condition of promotion, provide special tutorial help concurrent with promotion, etc.).

The limited nature of the options available concerning a pupil’s future at this level of the educational ladder argues for a simple system for recording and reporting teacher judgments. At most, it appears that three categories (e.g., Unsatisfactory, Acceptable, Outstanding) are sufficient for communicating to parents. Instead of worrying about more refined distinctions, elementary school teachers could better spend their energy working with individual pupils as they encounter learning problems. Teachers at this level are also in a position (because of self-contained classrooms) to spend a greater fraction of their time monitoring and reporting on the social and emotional development of their pupils. Such matters are clearly a special concern to parents of pupils at this level because problems in the social and emotional domain may have a direct bearing upon learning.

At the next educational level (grades 7-12), the options available to students are typically much more varied than at the earlier stage. The curriculum offers more choices both in terms of subject matter and in terms of special learning opportunities (e.g., accelerated courses, honors programs, work-study opportunities, vocational training, etc.). A student’s performance in school during this period plays a major role in determining possible postsecondary job options and/or opportunities for postsecondary education. Entry into higher education is especially significant because this is the gateway to those careers that are generally considered to be the most rewarding, both personally and financially.

The sorting of individuals during the 7th through 12th grades in U.S. education is extremely critical to individual students and to society as a whole. With rare exceptions, the educational choices made during this period of development will, for better or worse, have a profound impact upon opportunities later in life. There is likely to be a continuing debate over whether this is ultimately helpful or harmful to individuals and to society. Nevertheless, there is not a serious debate over whether this is, in fact, the current state of affairs.5

5Critics of traditional grading systems usually fail to recognize that the choices made by (and for) students will be made regardless of whether grades are available. Other sources of information (e.g., standardized test scores) will simply take on more significance as proxy indices of academic talent when grade data are either unavailable or nondiscriminating.
Given the educational system described, a somewhat more refined system for recording and reporting student grades than is used at the earlier levels seems desirable for grades 7-12. For example, a system with five categories (e.g., A, B, C, D, F/N) would provide sufficient differentiation so long as such a system is used properly. That is, there should be a reasonable spread of grades with relatively small frequencies at the extremes and proportionately larger frequencies in the middle category. This does not imply that the distribution should be "normal" (or even symmetric) in form. There are bound to be differences from class to class that justify different distribution shapes. However, it would be quite helpful if written schoolwide grading policies could be agreed upon that either suggest how grade distributions should look or place general constraints upon what individual teachers can do in assigning grades.

Generally speaking, the issues related to grading at the undergraduate level in college (grades 13-16) parallel those at the secondary level. Students in 4-year undergraduate programs still are faced with a variety of choices with respect to exploring new fields of study, choosing a major field of study, determining whether to pursue advanced study in graduate or professional school, etc. As previously noted, these decisions typically have a long-term impact upon an individual. From the point of view of a meritocratic social system, opportunities offered to students are afforded through a continuation of the sorting that begins at the secondary level. For reasons given above, recommendations concerning the nature of a grading system and how it should be employed in 4-year undergraduate programs are the same as for the secondary level.

There are other postsecondary educational settings where grading systems with fewer categories are appropriate. For example, vocational schools, trade/industrial training programs, and 2-year community college degree programs that are designed to prepare students for specific occupations share a common goal—providing students with the basic knowledge and practical skills necessary to succeed in a specific set of jobs. Here the primary question is, "Does the student possess the knowledge and skills required on the job?" Because the curriculum is ordinarily designed with the specific job demands in mind and students typically are provided with a substantial amount of job-like training as part of the curriculum, competency-based approaches to evaluation are highly appropriate. A grading system comprising no more than three categories will suffice under these conditions (e.g., lacks basic knowledge/skills [unsatisfactory],
possesses basic knowledge/skills [satisfactory], possesses knowledge/skills well beyond the basic level [exceptional]).

Students in graduate and professional school programs are comparable to students in vocationally oriented training in the sense that they are in the terminal stage of their education. Despite the fact that the knowledge base is broader and the cognitive outcomes are more complex at the graduate and professional school level, there seems to be little need for a highly refined grading system. After all, students at this level already have been subjected to extensive sorting and selection prior to their entry into the most advanced stage of their education. Given this fact, the expectation is that almost all students who are admitted will succeed. The major question is, “How well has this student performed in relation to others at the same stage of their education?” No more than three categories for recording judgments should be needed (e.g., unacceptable [U], satisfactory [S], outstanding [O]). Presumably, the first category would be employed rarely, the second category very frequently, and the third category with a fairly low frequency.

The Curriculum Factor

A second major variable that should be considered in discussing grades is the role played by a course in the overall curriculum. This is especially important beyond the elementary level because the curriculum becomes more diverse and student choices in selecting course experiences become more varied. For purposes of the present discussion, the curriculum can be partitioned into three major groupings: (a) core academic courses where outcomes are primarily cognitive in nature; (b) specialized courses in disciplines where the outcomes are defined in terms of self-expression in combination with psychomotor and/or affective processes; and (c) general elective courses that emphasize practical skills and/or psychomotor outcomes. The reason for making these distinctions is that performance in courses of different types has different implications for a student’s future.

Under the heading of core academic courses at the secondary level are classes in foreign languages, language arts (composition, literature, speech communications, etc.), mathematics (all types), science (biology, chemistry, earth sciences, and physics), social studies (civics, geography, history, etc.), and behavioral/social sciences (psychology, sociology, etc.). All such courses are core in the sense that they present foundational knowledge and concepts that provide a
framework for comprehending the world about us. These courses provide the building blocks for more advanced study at the postsecondary level. Therefore, valid information about how well students perform in these areas is especially crucial to making informed decisions concerning the likelihood of future academic success. Historically, grades have been the most valid indicator of future academic performance.

Specialized courses in which outcomes depend largely upon self-expression mixed with psychomotor and/or affective processes are those in the performing arts (dance, drama, musical [instrumental or vocal] performance, etc.), literary arts (writing of fiction and poetry), and visual arts (painting, lithography, sculpture, etc.). Courses of this type are different from core courses in that they rely heavily upon specialized and creative modes of self-expression. More importantly, they are different because they tap aptitudes that have, at best, a marginal relationship to future academic performance as defined by the core curriculum. Valid information about how well students perform in these specialized courses is likely to be useful primarily in predicting future success in the particular field of artistic expression.

There are special problems associated with evaluating artistic performances and creative works. For example, the judgmental standards employed are quite subjective and extremely difficult to define. It is frequently impossible to obtain a clear consensus among experts. To the extent that students are allowed individual discretion in creating performances and projects, there is a fundamental lack of comparability in the finished products. This frequently forces teachers to judge outcomes with respect to individualized expectations based upon beliefs that they hold concerning student talent. Some teachers in artistic fields refuse to make comparative judgments at all because they maintain that each creative work must be judged in terms of how well the artist achieved his/her own creative goals. All of these factors clearly suggest that grading in courses emphasizing artistic expression needs to be treated differently from that in core academic courses.

General elective courses that emphasize practical and/or psychomotor skills include vocational courses (distributive education, home economics, industrial education, etc.) and courses where outcomes relate directly to motor skills (physical education, keyboarding, shorthand, etc.). Obviously, these courses have a different function in the curriculum than do core academic courses. Some of these are designed to provide students with an opportunity to explore special interests and/or to develop practical skills useful in daily life. Others
are designed to give students a preliminary exposure to specific vocational activities. Still others afford an opportunity to engage in active athletic competition. It seems unrealistic to believe that performance in such courses has any predictive relationship to future academic success.

The Pitfalls of Generalization

The foregoing discussion of the factors that influence grading systems and grading practices underscores the folly of making sweeping recommendations concerning approaches to grading students. Grading methods that are appropriate under one set of circumstances may be highly inappropriate in another setting. Both the number of grading categories employed and the framework used by a teacher in judging performance need to be adapted to the educational context.

Unfortunately, there is a tendency to ignore important situational variables in discussions of grading. The implications of achievement in a core academic course in secondary school for a student's future opportunities are profoundly different than would be the achievement of the same student in home economics or physical education. Likewise, the outcomes of elementary school instruction have very different implications than do the outcomes in a required course for a first-year medical student.

The remainder of this chapter will focus upon grading in the core academic courses at the secondary and college levels. Based upon the premise that grading is an important, albeit distasteful, part of the job of teaching, general principles and specific guidelines for the assignment of grades at the secondary and college levels will be presented.

THE GRADE ASSIGNMENT PROCESS

General Principles

There are several general notions concerning grading that should be made explicit at the outset. Some of these ideas are rooted in philosophical beliefs, some come from a cognitive analysis of classroom learning, and others have their origins in classical measurement theory. All are important for teachers to understand if grades are to serve as a defensible basis for decision making.

---

6 Much of the material in this section is based upon a recent paper by the author (cf. Terwilliger, 1989).
1. Grading is a process of publicly certifying the teacher's judgment of the quality of a student's achievement in a specific course of study.

2. A teacher's judgment concerning student achievement should be based upon data that have been systematically collected specifically for that purpose. Only data that are directly related to achievement should be employed in grade assignment.

3. Grades should be assigned only as frequently as required by the school or college reporting system. This will allow for the collection of a sufficient amount of data to guarantee that grades are reliably assigned.

4. The assignment of a grade of "Failure" (F) or "No Credit" (N) has special importance. The basis for assigning such a grade should be a categorical judgment of the student's performance that is independent of the achievement of other students.

5. Realistic expectations concerning student achievement can only be obtained through experience. Teachers typically arrive at grading practices appropriate to specific settings through a process of trial and error.

The first principle is based upon the assumption that the meaning of a grade is clarified by considering only evidence directly linked to achievement. The utility of grades for decision making is diminished if a teacher attempts to factor in judgments of student effort, potential, work habits, etc. If the reporting system used requires the teacher to make such judgments, these should be recorded and reported separately from the grade.

Further, the quality of achievement in any subject matter should be defined in terms of the level of the outcomes achieved by students, not the amount of work students perform. There are several general hierarchical systems for defining the cognitive level of learning outcomes. Perhaps the best known is the Taxonomy of Educational Objectives (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). Another more recent system has been proposed by Presseisen (1986). These are useful for a variety of purposes beyond the assignment of grades.

The second principle assumes that grades are based upon some composite index derived from a clearly defined data base. This means

---

7 Contract grading schemes defined in terms of quantity of work or the granting of "extra credit" for work beyond that generally required should be discouraged. Such approaches may encourage and reward effort but they have no relationship to evaluation of quality.
that a teacher should have an a priori plan for collecting data. The amount of data collected should be sufficient to assure reliability and the variety of data should be sufficient to assure that the basis for judging achievement is broadly defined. No claim is made that this results in objective grades. However, it does make the grading process more explicit.

Although grades should be clearly linked to data, it does not follow that all data collected by a teacher need to be considered when assigning grades. There are other reasons for collecting data (e.g., giving periodic feedback to students, providing practice exercises, problems, quizzes, etc.) and obtaining data for the purpose of evaluating instruction, course materials, etc.

The third principle assumes a fundamental distinction between the process of judging performance and the process of data collection. It is well known that the validity of a judgment is enhanced if the data employed are reliable and relevant. Both reliability and relevance are improved when a substantial amount of data are collected over an extended period of time.

Critics of traditional grading are correct in saying that the importance of grades in the minds of students is frequently exaggerated. This is due, in part, to the inappropriate use of grades. The teacher who falls into the trap of assigning grades every time class assignments are due, quizzes are administered, projects are completed, tests are given, etc. is only contributing to many of the negative side effects of grades noted by Kirschenbaum et al. (1971) in Wad-ja-get? Teachers should learn to differentiate clearly between the act of making a judgment (assigning a grade) and the act of collecting data (obtaining information on which to base a judgment). Data collection should occur with much greater frequency than grading.

The fourth principle addresses the painful issue of failure or no credit. This is usually the "worst case" scenario for both a student and a teacher. The only way to avoid such a scenario is to refuse to consider a grade of F or N as an option. Some critics of grading endorse that approach. Whatever short-term benefit this has for the student may result in a long-term cost both to the student and to society (e.g., the student may subsequently be in a more advanced course or a job setting where unlearned knowledge and/or skills are critical).

The method for determining a grade of F or N should be as fair and honest as possible. Fair means that students know exactly what performance expectations define the boundary between F and "non-F." Honest means that the performance expectations are established
by the teacher based upon a thoughtful and thorough specification of the knowledge and/or skills that are regarded as minimal or essential outcomes of the course. A grade of F should result from the teacher's judgment that a student does not possess a minimal level of competence as defined by the essential course outcomes. In other words, failing grades should be assigned on the basis of a categorical (criterion-referenced) judgment rather than a comparative (norm-referenced) judgment.

The final principle is an acknowledgement that grading practices evolve with experience. Ideas regarding what data to collect, how to design assignments, tests, etc., and how to use the results have to be developed. Performance standards are established and modified in an iterative fashion. Norms, whether they be explicit or in the teacher's head, are built from long experience with different groups of students. In summary, developing a practical and valid set of grading practices is a long-term undertaking.

A Specific Approach to Grading

The five general principles discussed above provide a general framework for thinking about grade assignment but they do not provide specific guidance. This section will describe in detail an approach to grading that can be adapted to a wide variety of classroom settings. Prior to doing so, however, there are two specific recommendations that will improve grading practices regardless of the particular approach employed:

1. At the beginning of each term a teacher should prepare an evaluation outline for distribution to students. This outline should give dates for quizzes, exams, class presentations, etc. as well as due dates for assignments and projects. In addition, the outline should specify the nature of the quizzes and examinations (choice response vs. free response questions) and conditions under which they are to be administered (time limits, use of reference materials, etc.). Finally, the outline should clearly indicate the relative weight to be given to each item of data in arriving at grades.

2. All data to be employed in grading should be expressed in quantitative form. This implies that a teacher designs a scoring system, however primitive, for counting points earned on all quizzes, exams, assignments, projects, presentations, etc. The teacher should provide feedback to students in terms of
points earned rather than letter grades or some corresponding evaluation of performance.

Both of these recommendations are based upon the assumption that the teacher has acquired substantial experience with the subject matter in question. Therefore, these should be viewed as “end state” conditions after the teacher has experimented with different methods for collecting, coding, and aggregating data relevant to achievement in the subject matter.

**Minimal vs. developmental objectives.** In every subject matter and educational level, there are instructional outcomes that are essential or basic in the sense that they define the most rudimentary knowledge and skills. In principle, these are outcomes that every student is expected to achieve. Gronlund (1985) refers to such outcomes as **minimal objectives**.

In contrast to minimal objectives, in any subject matter and educational level there are also a large (and unspecified) number of instructional outcomes that define more complex and advanced levels of achievement. In principle, these more advanced outcomes are attainable only after students have mastered the minimal objectives. However, due to their diverse and subtle nature, it is *not* assumed that all students will achieve all (or even most) of them. Consequently, it is expected that there will be reliable individual differences among students with respect to performance on these more advanced outcomes. Gronlund (1985) has called these *developmental objectives* because they reflect a student’s level of development in striving to achieve the more challenging instructional outcomes.

The distinction between minimal and developmental objectives is crucial not only to the assignment of grades but also to designing instructional systems. For example, Gronlund (1973) argues that Bloom’s (1968) notions about mastery learning and mastery testing apply well to minimal objectives but are not as appropriate in the case of developmental objectives. The same distinction holds for all approaches to instruction (e.g., outcome-based or competency-based education) that emphasize all students achieving at the same a priori standard.

There are several ways to differentiate minimal from developmental objectives. For example, Gronlund (1973) defines minimal objectives in terms of the following questions:

1. What minimum knowledge and skills are prerequisite to further learning in the same area (e.g., knowledge of terms, measurement skills)?
2. What basic skills are prerequisite to learning in other areas (e.g., reading skills, computational skills, language skills)?

3. What minimum skill is needed for safe performance in some particular activity (e.g., using laboratory equipment, driving an automobile)?

4. What knowledge and skills are needed to attain minimum job proficiency (e.g., lathe operation, typing skill)?

5. What minimum knowledge and skills are needed to function in everyday, out-of-school, situations (e.g., reading, writing, speaking)? (p. 8)

Gronlund (1973) further suggests that the definition of outcomes that all students are expected to master be done cooperatively by teachers in consultation with subject matter authorities, curriculum specialists, and experts on learning.

A second way to distinguish minimal from developmental outcomes is to refer to cognitive analyses of instruction. Minimal objectives correspond to lower level cognitive outcomes, whereas developmental objectives correspond to higher level cognitive outcomes. For example, Presseisen (1986) describes four categories of thinking skills:

a. Essential cognitive processes—the basic thinking skills that are the building blocks of thought development;

b. Higher-order cognitive processes—the more complex thinking skills, which may be harder to define but which are based on the essential cognitive processes;

c. Metacognitive processes—the learning to learn skills aimed at making thinking more conscious and the student more aware of the ways one can go about problem solving or decision making; and

d. Epistemic cognitive processes—the kinds of thinking related to particular bodies of knowledge or subject matters and the particular problems addressed by these knowledge areas as well as the interdisciplinary relationships among content areas. (p. 9)

The first category might serve as a basis for defining minimal objectives, whereas some mixture of metacognitive and epistemic process could define developmental objectives. This is supported by Presseisen’s (1986) description of the difference between the first two categories:

There is a decided difference between what is meant as a higher-order thinking and the exact, standardized, minimal competency objectives often included in basic skills instruction. Simplistic, rote information that fits limited instructional sequences is not sufficient
as the material upon which to develop students’ higher-order thinking. (p. 11)

Terwilliger (1989) has argued that novelty is a useful basis for distinguishing minimal from developmental objectives:

One concept that I employ is novelty. I believe that outcomes that are defined as minimal objectives are those that test students’ ability to deal with familiar concepts and rehearsed skills. By definition, such outcomes have a low level of novelty. In contrast, developmental objectives test students’ ability to apply learning to new material or situations. (p. 17)

It can also be noted that the application of learning to new settings has historically been described as transfer of learning. Many years ago Thorndike (1969b) described this as the basis for teaching and testing for understanding:

The crucial indicator of a student’s understanding of a concept, a principle, or a procedure is that he is able to apply it in circumstances that are different from those under which it was taught. Transferability is the key feature of meaningful learning. So if we are to test for understanding, we must test in circumstances that are at least in part new.

Does a child really know how to read a map? Try him with one that is different from the one in the book. Does he really understand denominate numbers? Give him some problems phrased in “wugs,” “pogs,” and “pilzits,” the units used in measurement in the country of “Zoolumbia.” (I hope that a real “Zoolumbia” hasn’t sprung into existence recently without my being aware of it.) Does the Bill of Rights mean anything to him except a lot of words to be memorized? Ask him in what way recently proposed laws to regulate the sale of firearms might be considered unconstitutional. (p. 2)

**Minimal outcomes and failure.** A series of special quizzes, exercises, etc. should be designed to measure student achievement of the minimal objectives. These assessments function like mastery tests in the following ways:

1. Some a priori performance standard (for instance, 75% or 80% of maximum possible) is set for each assessment.
2. The expectation is that most, if not all, students (for instance, 90-95%) will perform at or above the level specified by the standard.
3. Students who fail to achieve at or above the standard will be given a second opportunity to take a parallel version of the quiz, exercise, etc. after review and remediation. The higher of
the two scores achieved (original vs. parallel version) will be recorded for the student.

It is important to compare the long-term failure rate on each minimal objectives assessment with the expectation that 90-95% of the students will achieve the minimal objectives. The statement of such an expectation provides a benchmark for determining if the difficulty level of the minimal objectives assessment is appropriate. Failure rates may be quite high for some quizzes, etc., suggesting that either they are too difficult or the standard is too high. For other measures failure rates may be zero, suggesting that either the learning outcomes are somewhat trivial or the standard is too low. It is important that the difficulty level of measures of minimal objectives outcomes be properly calibrated with the standard set for pass/fail decisions. This usually requires two or more administrations of a measure.

Warren (1971) has made the following insightful comment with regard to the setting of “absolute” standards:

Even in the British system of external examiners and in criterion-referenced testing, the “absolute” standard is established in relation to some expectation of performance based on past experience with examinees in similar circumstances. The real issue is in specifying the source of the standard on which grades are to be based. (p. 23)

An aggregate score on all minimal objectives assessments is determined for each student at the time grades are to reported. The score typically will be expressed as a percent of the maximum possible points on all minimal objectives assessments administered during the grading period. Pass/fail decisions should be made by comparing the aggregate score of each student to the a priori standard. Those who achieve the standard “pass” and those who do not “fail.”

Developmental objectives and passing grades. A separate set of achievement measures must be developed as a basis for differentiating levels of acceptable performance. These measures define differences among students in their achievement on the cognitively more complex developmental objectives. No a priori standard is specified for these measures. Instead, the performance of each student is interpreted with respect to norms derived from the administration of developmental objectives measures to reasonably large groups (for instance, 50 or more) of students. Normative data can be built up over time where class sizes are small.
4. SOME THOUGHTS ON GRADING

As previously noted, measures of developmental objectives should require students to apply knowledge and skills to novel settings. According to Fleming and Chambers (1983), this is not what teachers are accustomed to doing. Context-dependent questions that incorporate graphs, diagrams, tables, maps, etc. are useful devices for measuring cognitively complex outcomes. Teachers clearly need much more training than they currently receive in developing questions of this type. Teachers also need more practice in designing assignments, projects, term papers, etc. that require students to engage in critical analyses of novel situations, to integrate and synthesize familiar information with new data, to judge the merits of competing interpretations and contradictory evidence, etc. Activities such as these impress upon students the difference between low-level and high-level outcomes.

It is assumed that properly designed measures of higher order outcomes will result in score distributions in which the average score with respect to the maximum possible is much lower than for measures of minimal objectives. Also, the distribution of scores should be much more symmetrical in form with substantial variability. A summary of the expected statistical properties of the two types of measures is given in Table 3.

---

**Table 3. Expected Characteristics of Score Distributions Resulting From Minimal Objectives and Developmental Objectives Measures**

<table>
<thead>
<tr>
<th>Distribution Characteristic</th>
<th>Minimal Objectives</th>
<th>Developmental Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape</td>
<td>Definite negative skew</td>
<td>Approximately unimodal symmetric</td>
</tr>
<tr>
<td>Central tendency (difficulty level)</td>
<td>Mean score well above a priori standard (e.g., .05 to .10) when divided by maximum possible score</td>
<td>Mean score divided by maximum possible score in interval between .50 and .70</td>
</tr>
<tr>
<td>Variability</td>
<td>Can be small or large; depends primarily upon degree of skew in distribution</td>
<td>Should be quite large</td>
</tr>
</tbody>
</table>

---
Grades should be assigned on the basis of composite scores that combine data on several developmental objectives measures. (Presumably, the relative weight associated with each measure has been specified for students in the evaluation outline previously mentioned.) Assuming the composite score distribution is as expected, a norm-referenced basis for assigning grades can be employed readily. This is bound to be a trial-and-error process much like that for deciding standards for minimal objectives measures. However, with experience, teachers can develop very explicit norms that can be shared with students to help them understand the basis for judgments being made.

Two comments concerning norm-referenced grade assignment are in order. First, norm referencing does not imply a normal curve model. It is helpful if the distribution of composite scores is approximately unimodal and symmetric, but the main concern is that the variation is sufficiently great to assure reliable differences as the basis for grade assignment. Second, critics of grading often equate norm-referenced systems with direct competition among students. This is only the case when the norm group is restricted to others in the same class. The recommendation here is that the norms be based upon a more inclusive group (e.g., all students who have enrolled in the course over a specified period of time, for instance, during the most recent 3-5 years). This will result in much more stable norms and greatly reduce the competitive aspect of grades.

A summary of the grading process that has been described is presented in Figure 2. This makes it very clear that a two-track approach to evaluation is being proposed. One track leads to a dichotomous (pass/fail) decision employing a criterion-referenced model. The second track leads to a polychotomous (e.g., letter grade) decision employing a norm-referenced model. In courses where only pass/fail grades are used, the criterion-referenced model will suffice. In courses where students have the option of enrolling either on the pass/fail or traditional grading system, those on the pass/fail system are required to demonstrate achievement only at the minimal objectives level. Those enrolled on the traditional system must complete all assessment measures. For those students, grading is a two-stage process. First, students must demonstrate mastery of minimal objectives. Then, based upon performance on measures of developmental objectives, grades are assigned using norms.

Technical issues associated with weighting measures in the formation of composites are not discussed here. Terwilliger (1977) and Oosterhof (1987) provide detailed treatments of this topic.
Anyone who carefully examines the literature on grading systems and practices is struck by the continuing controversy over grades. Warren (1975) noted that recurring arguments over the purposes and definition of grades can be traced back to the period shortly after the turn of the century. Philosophical differences are at the heart of the controversy.

The scant attention given to the topic by authors of texts on classroom measurement tends to focus on practical and psychometric concerns. The recommendations given in these texts presuppose that teachers accept traditional grading as beneficial both to individual students and to society. However, there is substantial evidence that this is not the case.

Stiggins, Frisbie, and Griswold (1989) report that the majority of secondary teachers they studied employ grading practices that are at variance with the conventional wisdom offered in textbooks on mea-
surement. In discussing the research implications of their findings, Stiggins et al. (1989) refer to philosophical beliefs. They state:

It is a matter of educational values, for example, what information the grade assigned to a student should convey: achievement relative to others (norm referenced) or achievement relative to some absolute performance standards (criterion referenced). No research studies can help to answer the question, Which meaning should grades convey? A teacher’s judgment about the grading approach to be used should be dictated by the broader educational values (particularly the theory of teaching) that he or she holds. Until the teacher decides what meaning the grades should convey, most other decisions about grades and grading practices cannot be made. The significant research questions that need to be examined differ between these two grading approaches and even between methods within each approach. (p. 11)

The two-track approach to grade assignment that is recommended here attempts to demonstrate that teachers do not have to choose between criterion-referenced and norm-referenced approaches. Both can (and usually should) be employed in assigning grades. Each approach is uniquely suited to a particular problem faced in assigning grades.

At one level it is possible to obtain an empirical answer to the question concerning the merits of norm-referenced versus criterion-referenced grades. One simply has to compare the predictive power of grades assigned by the two approaches using criteria defined by subsequent performance in academic and/or employment settings. Of course, at a more fundamental level, the question of the relative costs and benefits of grades to students and society cannot be resolved by empirical research, no matter how sophisticated the methodology. Therefore, despite the virtues of any specific set of recommendations, it is safe to assume that the controversy over grades will continue.

REFERENCES


4. SOME THOUGHTS ON GRADING


Schofield, W. (1972). Memo to members of subcommittee on grading, University of Minnesota Assembly Committee on Educational Policy.


4. SOME THOUGHTS ON GRADING


Public education, and its assessment practices, have evolved from beliefs about how the world operates that are equivalent to the flat-earth theory. As long as the assessment of students is designed to fit obsolete “truths” about knowledge, learning, the mind, and human organizations, the roles and responsibilities of teachers in conducting such assessments, their purposes for doing so, and their preparation for fulfilling them will be likewise obsolete.

The following self-assessment (Figure 1) meets only one of fellow presenter H. D. Hoover’s criteria for tests. It has not been checked for validity, reliability, objectivity, or fairness. It is, however, feasible. Thus, I invite you to examine some of your beliefs. Indicate the degree to which you agree or disagree with each of these statements.

I start with this examination of beliefs because neuroscientists tell us that believing is seeing. Their work expands the observations originally made by Kuhn in 1962 (Kuhn, 1970) that the paradigms governing scientists’ work frequently prevent them from perceiving data that do not fit their particular structure of reality. What we believe about how the world works dictates what we are able to
We can understand things best when we break them down to their smallest component parts.

Learning consists of the sequential accumulation of discrete facts and skills.

Anarchy would reign if the staff and community of each school set their own goals for education and measured their attainment.

To be valid and reliable, assessment instruments should be developed by specialists.

Students should not be pressed to perform beyond their abilities.

perceive in that world. Our beliefs, attitudes, and assumptions comprise our world view—the lenses through which we perceive and understand the world. William Barrett, describing the futility of the philosopher who professed to make a really fresh start in his discipline, points out, “Alas, he is found to carry in his intellectual baggage assumptions unsuspected by himself, as a consequence of which his data became selective and screened. He cannot see the glasses through which he is seeing” (1987, p. 36). And, what has this to do with the title of this chapter?

Teachers’ lenses, or world view, inform their purpose in assessing students and that purpose dictates their roles and responsibilities. That purpose and its related roles and responsibilities are also focused by the world views of school administrators, boards of education, teacher educators, legislators, education department officials, researchers, test developers, and the public.

I intend to demonstrate how our beliefs define the current world of education and that corner of the education world we call “assessment.” And how a different set of beliefs—or paradigm lenses—could reveal an entirely different world.

I will share a story told by psychologist Jean Houston (1982) to illustrate what I mean by a world view and by my theme “believing is seeing.” She tells of a tribe in the Kalahari who believe that the world ends just beyond their local village boundaries.
It is reported that if you take them to that point, they will see nothing beyond it but a void. And if you tell them that you will prove to them that the world continues to exist beyond that point by stepping over the line of their so-called “world’s end,” they cry and beg you not to. If you persist in doing so, they are no longer able to see you and mourn your departure until you return across the line into the existing world. (p. 193)

This particular world view might be described as “magical.” Nevertheless, what these African tribesmen believe about their world literally dictates what they are able to perceive in it.

James Burke, narrator of the PBS series “The Day the Universe Changed” (1987), asserted that we all do what these tribesmen do. We “alter reality to make it fit what [we’ve] decided it should be. Without a structure, a theory, for what’s there, [we] don’t see anything. [We] have to have some version of reality . . . For things to make sense, [we] have to make [our] mind up about them in advance.”

Thus, our personal world views may constrain or expand our own sense of the boundaries of reality. Think for a minute about the kind of educational world that is created by the assumptions or theories that flow from an overarching belief that posits the nature of reality as mechanistic, reductionist, and deterministic. Component beliefs include the following:

- The mind is like a machine, taking in information from an objective, externalized environment, which functions on a linear time continuum.
- Intelligence is a static commodity which one either possesses or lacks.
- Learning occurs through the accretion of discrete, isolated bits of information and skills.
- The role of schools is to serve as a giant sieve for society, sorting and sifting its clients into their appropriate societal roles.
- Human enterprises operate most effectively when they are organized in a segmented, command-and-control hierarchy.

These beliefs are components of a world view and are illuminated by paradigms as Kuhn (1977) defines them:

Paradigms are not to be entirely equated with theories. Most fundamentally, they are accepted concrete examples of scientific achievement, actual problem solutions which scientists study with care and upon which they model their own work. If the notion of paradigm can be useful to the art historian, it will be pictures not
styles that serve as paradigms . . . . Both “style” and “theory” are terms used when describing a group of works which are recognizably similar. (p. 351)

Thus, paradigms in education appear to include the following:

- Horace Mann’s common school, consisting of a teacher-lecturer, supervised seatwork, and textbooks—all designed to control information.
- Charles Eliot’s 20th century curriculum consisting of new classifications of knowledge, clock hours, and Carnegie units.
- Conant’s consolidated schools, a combining of small, independent educational units into larger and segmented organizational hierarchies.
- Binet’s test of intellectual capacity.

The theories from which these paradigms emerged were “chosen” in a similar way to how scientific theories are “chosen” according to Kuhn (1977); that is, on the basis of both objective and subjective, or individually idiosyncratic, criteria or values. The latter flow from personality and biography, both forged in the cultural crucible of the individual’s time. The same crucible forges the so-called objective criteria. Overarching beliefs about how the world works and the values attributed to those beliefs are embedded in the cultural environment and give rise to the criteria by which information is judged, to the lenses through which information itself is perceived, and to the models or paradigms that are emblematic of aspects of those beliefs.

**HYPOTHETICAL MECHANISTIC-REDUCTIONIST SCHOOLS AND PRACTICES**

If we believe that the world is ultimately knowable by reducing it to its smallest components and that those components are the equivalents to the parts of a machine, what kind of schools and practices could we expect to emerge? We would probably see knowledge divided and subdivided into atomistic bits of information and skills, arranged by subjects that are kept separate by departmental structures, textbooks, allotted minutes per day, and closed classroom doors. The curriculum would be arranged like a string of pearls, described in detailed scope-and-sequence documents. Students would be classified by age categories and distinguishing labels such as gifted, learning disabled, emotionally handicapped, trainable
mentally handicapped, educable mentally handicapped, overachiever, underachiever.

Passage of students, or "raw materials," through such a system would be akin to the progress of a car down the assembly line, manufacture occurring bit by bit. Instruction would consist of "batch processing" via frontal teaching and direct instruction. Teaching effectiveness would be judged against a checklist of enumerated behaviors based on a standard teaching model in which teachers had been trained.

In such a system, teachers would be seen as skilled assembly line workers following orders from their foreman-principal who is following orders from central office administrators who are following orders from the board of directors which is complying with state and federal orders. These line-workers would perform their work in isolation, neither helping nor being helped by others, each order-taker resenting and blaming the perceived order-giver.

And, if the school was not operating up to par, what might we see? We would not be surprised to see new equipment and technology brought into the "plant" to increase efficiency. Nor would there be surprise in viewing the influx of additional personnel--more specialists to deal with increasingly refined categories of students, subdivisions of knowledge and educational programs; more supervisors to handle ever more specialized educational functions and to oversee instructors who they see as Henry Ford saw his "average worker, want[ing] a job in which he does not have to think" (Clancy, 1989, p. 196).

**HYPOTHETICAL MECHANISTIC-REDUCTIONIST ASSESSMENT**

And what of assessment in a school that looks like this--a school that is designed to fulfill theories about intelligence as static, learning as simple accretion, the mind as machine, organizations as command-and-control hierarchies? It would undoubtedly be designed to fulfill the school's belief that its purpose is to prepare students for adult life by sorting and grading and labeling them as they are processed from raw material to finished product. A hard-headed "scientific" approach to testing would be employed that promised prediction and control. To legitimate its sorting process, the school would become increasingly dependent on the "certainty" that the mathematics of statistics brought to its judgments about students. As Barrett (1987) reminds us, "there is a certain type of mind that prefers exactness, or what looks like exactness, to adequacy" (p. 44).
We would probably see written and oral questions that seek the recall of atomistic bits of information and tests that check the acquisition of discrete skills. Formal and informal assessments would be devised to pit student against student so that their learning could be compared on a competitive basis, educational gates could be opened or closed to them, a rank-ordered value could be ascribed to each one, and their identification as "above average" or "below average" could be ascertained.

As public schools became more segmented and specialized and the desire to categorize students became more paramount, experts would be more heavily relied upon to devise quality control assessments that checked how many pearls were strung by each student as well as the strength of the student's string itself. These experts would perform their work for textbook publishers, state departments of education, school district central offices, universities, and commercial test companies. Cost containment concerns could be expected to limit the format of their expertly constructed assessments to a fill-in-the-bubble, machine-scorable one. Of course, the use of technology by such experts could allow them to establish data banks of test questions that might be accessed in a somewhat wider variety of formats by teachers. The common denominator of all such tests would be the presence of only one right answer for any item.

In such a mechanistic, reductionist world of education and assessment, the teachers' purpose in assessing students would be to determine the place for each student and then keep him/her in it. Their roles would be those of middle-man--delivering others' assessments to their students and returning completed forms to others for scoring and interpretation--mimics of their own teachers' and professors' assessment behaviors, and deterministic prophets of their students' success in school, based in large part on the information they gather while playing these other two roles. Their responsibility? To obey the directions of their "betters," their supervisors, the authors of teacher-proof textbooks, and assessment experts.

REAL SCHOOLS

Real schools in America bear a remarkable resemblance to those that were hypothetically modeled on the deterministic, mechanistic, reductionist world view of classical science. These real schools serve age-segmented clients who, for the most part, are treated as passive recipients of "lockstep applications of skill hierarchies and spiraled curriculum" (Marzano, 1988, p. 17). Identified at ever earlier ages as
“at-risk” or “gifted” or “developmentally slow” or as whatever the educational label-of-the-month might be, students are measured for their “fit” with the common curriculum and slotted into categories of bureaucratic convenience. “Where formerly classroom teachers tried to accommodate students’ diverse needs, . . . [they now] simply seek referrals to special programs for those students who do not fit the curriculum” (Keating & Oakes, 1988, p. 10).

And, regardless of which category they are in, most students can expect the mind-numbing, repetitive, educational rite of passage embodied in drill-and-practice of discrete, decontextualized subskills in a progression from easy to difficult—practicing over and over and over again what they do not master quickly and over again what they do. “[T]he curriculum and instructional strategies that are common across all tracks are often mediocre even for average and above-average students . . . . Telling and lecturing, along with monitoring seat work, dominate classroom teaching” (Keating & Oakes, 1988, p. 9).

Although today’s instructional practices have their roots in the lecture and seat work methods devised by Horace Mann in the 1830s to meet Massachusetts industrialists’ needs for a compliant work force, the “modern” high school curriculum, towards which the elementary and middle school curricula now spiral, was defined at the turn of the century by NEA’s Committee of Ten, chaired by Harvard President Charles Eliot. Arguing at the time that the 19th century curriculum of Latin, classical literature, rhetoric, natural philosophy, and natural history was irrelevant to the 20th century, this group recommended instead 4 years of English, 3 of social science, and 2 each of mathematics, science, and foreign language (Hutchins, 1988). Mimicking the high school division of knowledge into discrete categories, the elementary classroom lacks only the Pavlovian bell and its teachers the department chair status and extra pay of their secondary counterparts. Young students who enjoy playing “teacher,” like I did as a child, can duplicate their own teacher’s schedule of lessons with great accuracy after a few weeks in a particular classroom: each subject treated separately in sequence for the same number of minutes each day with occasional variations for science, art, music, and physical education.

Breaking curriculum, seen as classroom routines, into small steps and teaching the steps, and managing smooth transitions from one subject to another are two hallmarks of what has come to be known as “effective teaching.” Additional elements of this paradigm, which is based on research with a scope that applies primarily to elementary grade students in low SES schools for their acquisition of basic skills
as measured by standardized tests, include direct instruction of the entire class and assigning tasks with simple unambiguous demands (Edelsky & Harman, 1987).

Madeline Hunter epitomizes this behavioristic, cause-effect approach to teaching and personifies teacher-as-authority-figure, as well as conveying submission to the mega-authority of science, when she stands before her educator audiences, clad in a white lab coat while presenting her seven-step method of teaching. Although she calls her formula a model for teacher decision making and claims that it is not intended for summative evaluation of teacher performance, 44 out of 48 states surveyed by Democratic Schools in 1988 reported using Hunter and Hunter-type models as a teacher evaluation tool (DiBernardo & Stiles, 1988). Having been exposed to the Essential Elements of Instruction (EEI) through inservice sessions that are frequently mandatory, teachers are expected to use immediately every element in every lesson. Even when teachers’ instruction in EEI is accompanied by technical coaching, that process itself “fits excellently into an educational system which is becoming ever more inclined to bureaucratic forms of control over its employees in order to secure the implementation of centrally determined, standardized forms of ‘effective’ instruction” (Hargraves & Dawe, 1989). The standardized curriculum and standardized school day are now joined by standardized teaching practice--practice that is not linked by research to improved student learning (Slavin, 1987).

Thus “Hunterized,” teachers return to isolated classrooms, which comprise the lowest level of the educational hierarchy, once again having been the captive recipients of an expert’s knowledge and once again themselves becoming the expert dispensers of knowledge to their captive recipients.

The hierarchical, inflexible, top-down management structure of the entire educational system resembles a set of nested boxes. As policy makers at the federal and state levels mandate educational programs, procedures, and now, goals and the tests by which to measure them, school district decision makers, modeling their enterprise on Industrial Age corporate structures, mimic the contextual hierarchy in their organizations. Schools, like little factories, are characterized by lock-step learning, chopped up in discrete blocks of time and narrow notions of performance, and, like factories, turn out recognizable similar “products” over time (Keating & Oakes, 1988). Within the next nested box, teachers mechanically replicate the authoritarian hierarchy. Hutchins (1987) sees this phenomenon related to rigid evaluation models of teaching.
...the trend to mechanistic models of teaching seems to have been accompanied by a subtle shift in the responsibility for learning from the student to the teacher. As teaching becomes more and more a matter of “step 1, step 2, step 3, etc.,” the instructor becomes more and more controlling in the learning situation. The practice of using models tends to degenerate easily into saying, in effect, “Students, just follow my instructions and you will learn what I want you to learn.” (pp. 17-18)

Physically separated from their co-workers by the walls and closed doors of classrooms and professionally separated from the wisdom of their peers by sink-or-swim induction and operational norms, teachers try to establish a beachhead of control with their students. In such an environment, “where there are only crumbs to share,” Ann Lieberman (1988) finds that “teachers tend to hide their successes as well as their failures. Each teacher looks out for his or her own welfare . . . [amid] the powerful infantilizing effects of the school.” (p. 651)

In this “sage-and-fools caste system . . . [teachers’] present roles as classroom masters are more like wardens, more concerned with keeping their charges in line than with enabling them to live more fully” (Litvak & Senzee, 1986, p. 176). And, within such a system, the last nested hierarchical box is the brain itself, where educational practices lead to a dominance of the logical, analytical, step-by-step modes associated with the left hemisphere of the brain over the integrative, synthetical, and holistic modes of the right (Russell, 1983).

REAL ASSESSMENT

The actual assessment of students in our schools today resembles in almost every detail the hypothetical approaches proposed as derivatives of the mechanistic-reductionistic world view. Whether it is conducted formally or informally, student assessment is congruent with the educational structures and practices found in our schools.

Formal assessment of student achievement through the use of tests that were developed external to the school took root after World War I (Ornstein & Erlich, 1989), coinciding with the movement at the end of the 19th century, observed by Timar and Kirp (1988), away from an appreciation of education for its intrinsic value towards an appreciation of it for what it could do, its instrumental value.

At various times over the past eighty years, education has been regarded as creating social and political harmony by integrating
immigrants into the mainstream of American life, as creating a more “efficient” society ordered along industrial forms, . . . and as helping to regain America’s competitive edge in international economic markets. (Timar & Kirp, 1988, p. 46)

The emergent pragmatism of education found a happy marriage with the utility of the large-group administered and easily scored tests devised by Alfred Binet to determine the capacities of different children for schooling and modified for use by the Army in World War I as a quick measure of soldiers’ capabilities. Their scores on such tests became the primary determinant of their assignments. “Group administration of tests (a wartime necessity) and the strong reliance on the test score as the measure of ‘intelligence’ remained the norm even after the war ended” (Ornstein & Erlich, 1989, p. 109).

That intelligence and, later, achievement were thought to be reducible to a numerical score helped serve the instrumental purpose of education—one’s “number” determines one’s appropriate place in the clockwork universe—while reaffirming the belief that complex phenomena, such as human potential, could be understood best by breaking it down to its basic building blocks, an intelligence quotient, or an achievement score. Predicting academic success, the original purpose of such tests, has tended to determine academic success. As Keating and Oakes (1988) point out:

> Popular views about intelligence and ability, as well as perceptions about the distribution of talent in the general population, influence educational practice. What seems fair and reasonable at the moment—tests showing how students compare with others on global characteristics such as mathematics and verbal aptitude—turns out systematically to limit some students’ access to knowledge. For the most part, tests of intelligence, ability and achievement simply rank students, separating and segregating them and sorting them for future social participation . . . . Once the tests identify and legitimize students’ differences, students are provided with different school experiences. (p. 7)

In addition to their predictive uses, assessments of student achievement in the form of norm-referenced tests, criterion-referenced tests, and minimum-competency tests are also being used to gauge the success of schools themselves. In his presentation at the 1985 ETS Invitational Conference, Theodore Sizer attributed this accountability drive to “The public (or, more accurately, that minority of the public that has political awareness and clout) want[ing] to see evidence that its educational investment yields demonstrable returns” (p. 2).

Formal assessment of student achievement is a growth industry. The National Governors’ Association (1988) reports that in 1985 alone,
27 states adopted 37 new testing programs, supporting the contention of Dorr-Bremme and Herman (1986) that "educational testing is a pervasive enterprise . . . in which hundreds of millions of dollars in public monies are expended annually [and in which] significant teacher and student time is spent, representing fully half of the testing at the elementary school level and one-quarter of the total student testing time at the secondary level" (pp. 2, 18). Calfee and Hiebert (1987) decry the absence of programs to enhance teachers' skills in assessment as parallels to the burgeoning assessment programs mandated nationally and at the state level, concluding that because the state and national data bases of these programs make no provision for the judgments of classroom teachers, classroom assessments are not viewed as a sound base for policy making. Former Education Editor of the New York Times, Fred Hechinger (1989), opines that policy makers lack confidence in teachers. "If you trusted the teacher, you would say, 'This teacher can tell me how well this child does.' Since we don’t have that trust, we superimpose the tests" (Hechinger, 1989, p. 4).

Although policy makers seem to distrust teachers' assessments of student achievement, perhaps believing that they lack objectivity—a quality greatly valued by those operating within the mechanistic-reductionist world view, they appear to be unquestioning consumers of "standardized tests [that] are consistently sold as scientifically developed instruments that objectively, simply, and reliably measure students' achievement, abilities, or skills" (Neill & Medina, 1989, p. 689). Teachers themselves, although critical of standardized tests (Dorr-Bremme and Herman [1986]) in their 5-year study of test use found that teachers believed the tests were not a good measure of what they had taught and that they had a better, more specific idea of students' strengths and weaknesses), "proceed to test in predictable ways, often modeling their approaches on the externally developed examinations they see most often, the standardized achievement test. Or they simply use the tests included in the textbooks" (Atkin, Patrick, & Kennedy, 1989, p. 76).

**PREPARATION OF TEACHERS**

This conference poses the question, "Are our school teachers adequately trained in measurement and assessment skills." Given the current structure of our schools and the beliefs on which they are based, I must answer "yes" to this query. Teachers are exposed to little or no information on measurement and assessment in their
preparation because they are not the people in the education system who are expected or trusted to perform valued measurement or assessment. They receive as much training in this as the system will allow them to actually use.

Other chapters in this volume suggest that few teacher education programs require or offer coursework in student assessment. There seems to be little need to do so if it is experts within or outside of the public school system upon whom we are going to rely to perform the only valued performance of this function. I interpret the absence of such training as a revelation of the attitude that “Whatever you do to assess students in your classroom is okay, because it does not really count anyway.”

We have been told that when college coursework does include measurement and assessment, it tends to concentrate on statistics as a form of esoterica, knowledge to be grasped by only the chosen few who are far removed from the hurly-burly of the public school classroom. I interpret this training emphasis as manifesting an attitude that says in effect, “We, the Ed. Psych. gurus of tests and measurement, know what’s best for you. Because most of you won’t even fathom this, please trust us and our fellow experts to provide you the only credible assessment tools you’ll need once you reach the classroom. Go forth to sort and label, delivering our tests, imitating our guru-like demeanor, and following our directions.” Calfee and Hiebert (1987) describe this role of teachers, for which they are groomed by preservice preparation and the school workplace, as “meter reader” (p. 45).

THE OLD LENSES

Our schools and attendant assessment practices “make sense” when seen as grounded in the mechanistic-reductionist world view. It is our beliefs that provide versions of reality, James Burke instructs us:

For things to make sense, you have to make up your mind about them in advance; otherwise you wouldn’t know where you are. . . . The only structure in the shifting, changing face of nature is the one we impose on it with our theories, each one the latest version of what we call the truth. (Burke, 1987)

If we believe that the world ends here, then that is where we see the end of the world. If we believe that the nature of the world is analogous to a machine and can be understood when broken down to
its smallest component parts, then we derive understandings about the mind, intelligence, learning, and organizations that are consistent with that world view, and create enterprises that fit it. And, we see what we believe.

Physicist Fritjo Capra (1982) describes the traditional scientific view of the world, the one on which our schools are based, in this way:

Matter was thought to be the basis of all existence and the material world was seen as a multitude of separate objects assembled into a huge machine . . . . Consequently, it was believed that complex phenomena could always be understood by reducing them to their basic building blocks . . . . This attitude, known as “reductionist,” has become so deeply ingrained in our culture that it has often been identified with the scientific method. The other sciences accepted the mechanistic and reductionist views of classical physics as the correct description of reality and modeled their own theories accordingly. Whenever psychologists, sociologists, economists wanted to be scientific, they naturally turned toward the basic concepts of Newtonian physics. (p. 23)

Bela Banathy, general systems scientist, invites educators to consider the traditional scientific paradigms, dating from the 17th century, as we attempt to redesign the educational system:

Inspired by the Cartesian-Newtonian scientific world view, disciplined inquiry during the last three hundred years sought understanding by taking things apart, seeking the “ultimate” part, and groping to see the whole by viewing the characteristics of its parts. Implicit in this approach is an exclusive commitment to defining elementary cause and effect relationships, which led to a deterministic perception of the world. The outcome of these perspectives was best manifested in the Industrial Revolution, and its essential characteristics were derived from analytic thinking, reductionism, and determinism. (Banathy, 1988, p. 52)

Comfort with the belief that one knows (or can know) what causes things to happen and that the same conditions always produce the same results is typical of Second Wave thinking, Toffler (1981) tells us, and conjures up an image of the entire universe as consisting of “cue sticks and billiard balls--causes and effects” (pp. 303-304).

Embedded in the reductionistic, mechanical, and deterministic components of the Newtonian-Cartesian world view, is the related belief that change is incremental, occurring linearly. Believing thus, how could we see the mind as anything but a machine, learning as anything but cumulative, assessment as anything but a sorting and labeling process, and schools as anything but segmented hierarchies?
It was adherence to this Newtonian world view that guided the work of psychologists in the first half of this century which, according to Hampel and Farnham-Diggory, forms the basis of our present school system.

The ideas of Thorndike, Skinner, Gagne, Bloom and others shaped the organization of the school day, curricular materials, grading practices, and testing . . . . Thorndike and other viewed knowledge as modestly analogous to a string of pearls. Learning was the activity of stringing the pearls . . . . Curricula today are still largely based on the assumption that knowledge can be added to previous knowledge in a purely cumulative fashion. This is particularly evident wherever teachers are required to set so-called behavioral objectives for their pupils. Any plausible objective will do, as long as it can be counted, as long as students can spell eight out of ten words on list A. (Hampel & Farnham-Diggory, 1987, pp. 7-9)

Schools seem to have embraced the same “overly reductionistic, materialistic, and mechanistic” old-physics belief system that Litvak and Senzee (1986) accuse biology of emulating by “attempting to reduce biological phenomena to elementary bits and pieces. Many biologists today do not consider a biological phenomenon real unless it is reducible to an explanation from physics. Thus everything in the living world is reduced to machinery—all living things are ‘nothing but’ passive automata manipulated by the environment” (pp. 48-49). The irony is that the very world-view lenses of the old physics that are now worn by most biologists and educators, “reductionism based upon the mechanistic model of physics” (p. 49), has actually been rejected by the physicists themselves.

THE NEW LENSES

Just as earlier beliefs, “versions of the truth” (e.g., the earth is the center of the universe, the earth is flat, evil exists in the form of witches and burning them at the stake is an act of mercy, man is not meant to fly, children with Down’s Syndrome should routinely be institutionalized), worked perfectly well for a while, they eventually gave way to a new structure of reality. Kuhn (1970) describes this “giving way” of the paradigms governing science as following a predictable sequence: Prevailing images encountered anomalies. Uncertainty paved the way for competing images. Competition among paradigms held sway until one prevailed.

Jarman and Land (1989) summarize a paradigm shift in science that continues to reverberate:
5. TEACHERS' ASSESSMENT OF STUDENTS

Soon after the turn of the century, scientists ran smack into a collection of discoveries and facts about nature that forced them to totally revise their definition of reality . . . . Einstein’s discoveries completely redefined reality. [His and] a host of other pioneering ideas confirmed that the real world is based on entirely different principles than had been known or even suspected before this century. We are only just now beginning to recognize the almost incredible impact of those discoveries . . . . No one was offended more than these pioneering scientists themselves when their own discoveries and tests showed that the great body of ancient, logical and reasonable ideas of science was in error. The logical “natural order of things,” long thought to be the basis of nature, just did not fit the torrent of emerging facts. (pp. 39, 44)

Einstein showed that time and space were not absolute and fixed, but relative. Not only were time and space one but so were the electric and magnetic forces, and energy and matter (Russell, 1983). The discovery that “matter” is in fact bound energy revealed that everything in the universe exists in two very different and simultaneous states, as both particles and waves, as both something solid and invisible at the same time. “The world and everything in it, exists in two simultaneous and factual states: ‘being’--the physical, material state--and ‘becoming’--the invisible waves of possibility and probability surrounding it” (Jarman & Land, 1989, pp. 47, 50).

Einstein’s Theory of Relativity was followed by Quantum Theory: the behavior of subatomic particles appears random in nature. Einstein could not accept this paradigm shift, clinging to his deterministic lenses when he exclaimed, “God does not play dice with the universe!” and believed he had proved its discoverer, Max Planck, wrong with his Einstein-Podolsky-Rosen Effect (EPR Effect). However, nearly 50 years later, physicist J. S. Bell validated quantum mechanics with a test based on the EPR challenge and discovered that change in one particle which was smaller than an atom and moving at a velocity near the speed of light, simultaneously affected the other particle with which it had been paired. Its far-reaching implication: Everything in the universe is intimately connected without regard for the distance between any two objects (Travis & Callendar, 1990, p. 55).

Might it be this new version of the truth that provides the context for Heisenberg’s Uncertainty Principle? It contends “that it is impossible to objectively measure anything [because] the measuring device always interferes by forming a relationship with the subject that alters how the event in question would have turned out if no measurement had been taken” (Travis & Callendar, 1990, p. B-6). Heisenberg’s demonstration that the act of observation itself affects that which is
being observed had shattering implications for physicists of the time who regarded the observer and the observed as separate detached entities. "Somehow the mental and physical worlds were interdependent" (Russell, 1983, p. 141).

British physicist David Bohm offers one approach to understanding this interconnectedness with his notion of implicate order, or enfolded order, from which the explicate order, the universe we see around us, unfolds and into which it enfolds, simultaneously. The image of a hologram serves as a metaphor for implicate order. When even one part of a holographic plate is illuminated, an image of the whole object is still obtained (Bohm & Peat, 1987). This analogy suggests that each part of the physical universe—you, I, a tree, etc.—like the hologram, has the whole of time and space encoded in every part of it, containing all the information about the whole universe within it.

Jarman and Land (1992) conclude from these discoveries that "the ancient notion that all things are separate is factually wrong. Everybody and everything is connected. Everything affects everything else. No matter how different, no matter how far away, we are all part of one another" (p. 56).

Although the paradigms of relativity and quantum mechanics undermined the old paradigms of mechanical materialism and reductionistic separateness, their assumptions of randomness are being challenged by the theory of dissipative structures posited by Ilya Prigogine and by emergent chaos theory. Prigogine won the 1977 Nobel prize in chemistry for his study of the transformation of randomness into order, or the emergence of order from chaos. His Theory of Dissipative Structures proposes that inherent in the nature of any system is its attempt to stabilize itself in the midst of stress from the outside. If the stress becomes too great, the system may collapse. Alternatively, if the system survives this period of chaos, reorganization at a higher level of complexity and a new level of stability can emerge. Furthermore, the new is totally unpredictable if all we look at is the structure of the old (Travis & Callendar, 1990).

The past, thus, does not predict or cause the future, nor does constant change point to the devolution of molecular disorder. Rather, "Change is driven by the pull of the future to connect everything at broader, deeper, more interpenetrating levels . . . . Our world is progressing inevitably toward more complex interrelatedness and connectedness" (Jarman & Land, 1992, pp. 60-61).

Chronicler of chaos theory, James Gleick, reports on the scientific community's latest revolution which deals with the concept that from
seemingly chaotic behavior, regular but unpredictable patterns emerge. “John Hubbard . . . considered chaos a poor name for his work, because it implied randomness. To him, the overriding message was that simple processes in nature could produce magnificent edifices of complexity without randomness” (Gleick, 1987, p. 306). Furthermore, although the original investigators of chaos, who came from multiple scientific and mathematical specialties, expected their studies to support their tacit beliefs about complexity—that simple systems behave in simple ways, that complex behavior implies complex causes, and that different systems behave differently—they learned instead: “Simple systems give rise to complex behavior. Complex systems give rise to simple behavior. And most important, the laws of complexity hold universally, caring not at all for the details of a systems’ constituent atoms” (Gleick, 1987, p. 304).

As the new science of chaos itself arose from simultaneous inquiries by scientists in the fields of meteorology, mathematics, biology, physics, and astronomy, more and more of the investigators “felt the compartmentalization of science as an impediment to their work. More and more felt the futility of studying parts in isolation from the whole. For them, chaos was the end of the reductionist program in science” (Gleick, 1987, p. 304). Likewise, it marked the end of the either/or thinking of determinism or free will and the beginning of a marriage that wed determinism and free will (Gleick, 1987, p. 304).

WHEN WORLD VIEWS COLLIDE

A comparison of the basic beliefs comprising classical science with those of new science is shown in Figure 2.

Figure 2.

<table>
<thead>
<tr>
<th>New Science</th>
<th>Classical Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creative World View</td>
<td>Causal World View</td>
</tr>
</tbody>
</table>

- Change is probabilistic, occurring through a perpetual creative and transforming process of being and becoming.
- Change is driven by the pull of the future to connect everything at broader, deeper, more interpenetrating levels.
- All things are connected—at all times and instantaneously at any distance.
- All things are separate, existing independent of each other and their environment.
- Events are driven by and are a result of past causes; the present is determined by the past.

(Jarmin & Land, 1992, pp. 37-65)
The emerging paradigm in “physics, psychology, and progressive biological circles,” according to Litvak and Senzee (1986) “is that a comprehensive (or holistic) perspective must be adopted in place of mere mechanistic reductionism” (p. 49). Banathy (1988) characterizes the new scientific paradigm as “a major shift toward synthesis, expansionism, indeterminism, emergence, and a systemic-ecological world view” (p. 53). And Toffler (1981) describes the collision of this new world view with the entrenched paradigms of industrial society as “the beginnings of a philosophical revolt aimed at overthrowing the reigning assumptions of the past 300 years” (p. 289).

Kuhn (1970) calls such paradigm-induced changes in scientific perception “transformations of vision” (p. 118). A paradigm shift is not gradual, but a Gestalt shift--one must see it one way or the other. Proofs and logic are not the currency of exchange between conflicting paradigms. The irrelevance of one another’s arguments to adherents of competing paradigms has been labeled by Kuhn as “incommensurability.” “Communication across the revolutionary divide is inevitably partial. . . . before they can hope to communicate fully, one group or the other must experience the conversion [emphasis added] that we have been calling a paradigm shift . . . . The transfer of allegiance from paradigm to paradigm is a conversion experience that cannot be forced” (Kuhn, 1970, pp. 148-151).

Max Planck (as quoted by Kuhn, 1970) reflected on this in his autobiography:

A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it. (p. 151)

**BLIND SPOTS**

Why don’t scientists see data that they don’t expect to see? Why don’t many of us in education see the relationship of scientific breakthroughs to our own enterprise or see that our current paradigms limit human growth? I propose that the world-view lenses through which we perceive are by their very nature equipped with a blind spot that is analogous to the physical blind spot in the eye itself. Ornstein and Erlich (1989) point out that there are no photoreceptor cells where the optic nerve exits the retina, so this part of the retina cannot respond to light. “When the lens of the eye focuses an image of a small object on the blind spot, the image disappears. We don’t notice the loss: our brain simply fills it in, using the context of the rest
of the picture. It is not that something that was there is now gone; we don’t know that anything is missing” (p. 71).

Like the nested boxes of educational hierarchies, our blind spots are nested within one another. The world “out there” that we perceive is actually within us—in the design of the senses, the wiring of the nerve circuits, the processing of information in the brain, the interpreting of information (Ornstein & Erlich, 1989). So our physical blind spot is nested within an information-processing blind spot that is nested within an interpretive blind spot of attitudes, beliefs, paradigms, and world view. We don’t see what we don’t see. And, we don’t see what we don’t believe is worth seeing or is present to be seen.

“One particularly significant mechanism which the mind employs to defend itself against the inadequacy of its basic ideas,” according to Bohm and Peat (1987), “is to deny that it is relevant to explore these ideas” or to go even further and to deny implicitly “that anything important is being denied” (pp. 22-23). Likewise, we don’t see what we don’t believe is related to our particular area of inquiry.

Another way of defending the subliminal structure of ideas is to overemphasize the separation between a particular problem and other areas . . . . But this only acts to prevent a clear awareness of the ultimate connections of the problem to its wider context and implications. The result is to produce artificial and excessively sharp divisions between different problems and to obscure their connections to wider fields. . . . by ignoring the connections of each thing to its whole context, the illusion can be created that the ideas, structures, and institutions that are the dearest can go on indefinitely and unchanged. (Bohm & Peat, 1987, pp. 23, 208)

As long as we peer through the lens of “all things, including people, are separate and apart,” we will, for instance, deny that students and teachers should be dealt with as anything different from separate “particles” aligned within the hierarchy of a closed system. We will frame our problem as being one of how to better align those particles. Through those lenses we will also see schools as separate from their environment, thus seeking improvement efforts within the confines of the educational organization alone. Timar and Kirp (1988) describe this view as regarding “educational excellence as a series of discrete problems to be solved. The perspective is not broad . . . but narrow—on how to effect changes in specific areas of institutional life” (p. 120).

The more we argue for our blind spot, of course not knowing that it is blind, the blinder we become. “Rigidity,” say Bohm and Peat (1987), “is ultimately the very source of this deterioration . . . because
all the proposed remedies are actually different forms of the same illness that they aim to cure" (p. 209).

But, vision is transformed. World view lenses do change. Paradigms do shift.

**PRESCRIPTION LENSES INSTEAD OF WAR**

Kuhn, Bohm, Peat, and Russell give us clues regarding the conditions necessary for improved sight. For instance, Kuhn (1970) describes crisis as a necessary precondition for the emergence of novel theories. He also suggests that, although proponents of different theories are like native speakers of different languages whose communication problems are compounded by their frequent use of the same vocabulary to represent completely different concepts, they can attempt to exhibit to one another the concrete technical results achievable by those who practice within each theory (Kuhn, 1977).

Bohm and Peat (1987) likewise hold out hope for something less destructive than a revolution to enable a change of world-view lenses to occur. They argue for allowing a plurality of basic concepts, with a constant movement of free creative play that is aimed at establishing unity between them. Exercising the creative intelligence that perceives new categories and new orders “between” the older ones—disjointed extremes—calls for, in their opinion, (a) self-awareness—revealing one’s rigid assumptions to one’s self, and (b) “dialogue”—communicating with an open mind and an open heart, desirous of understanding the other’s point of view, ready to acknowledge any fact and any point of view as it actually is, and ready to change one’s own point of view if there is a good reason to do so.

Altering one’s world-view lenses may also consist of deliberately changing one’s mind set, suggests Russell. He reports on the work of Dutch futurist Fred Polak, which reveals that in every instance of a flowering culture there has been a positive image of the future at work and that the intensity and energy of the images have been reliable predictors of the direction that cultures would take. Russell quotes Polak’s conclusion: “Bold visionary thinking is in itself the prerequisite for effective social change” (Russell, 1983, p. 223).

**IMPLICATIONS FOR EDUCATION’S PARADIGMS**

Our education system, which has its roots in the Newtonian universe and its offspring, the Industrial Age, and which has survived for over 150 years by only modestly refining its essential components,
is under attack. Beleaguered by politicians, business leaders, and citizens, beset with internecine warfare, benumbed and benumbing with classroom routines, schools seem to be encountering both the crises that precede the appearance of a novel theory and the rising disorder that is prelude to a shift in order--to either collapse or to a higher level of interrelated complexity.

Ainsworth-Land and Ainsworth-Land (1982) describe this process of evolutionary creativity, of divergence and convergence, in *Forward to Basics*:

In any system, once a relative orderliness has been achieved, the only means by which a broader and more complex interrelationship among the various elements can be achieved is by introducing or generating disorder. The system can come apart to be put together in a much more integrated way. Any system that resists this creative disintegration and re-integration can only suffer the gradual erosion of its established order due to the energy required to protect the system from change. (p. 79)

Banathy (1988) sees the efforts to change and improve education during the last two to three decades, because they flow from the piecemeal, fragmented, so-called scientific approach, as having the effect of protecting the system from change.

Nevertheless, as scientific thought is transformed by pioneers who ask novel questions and see anew, so is educational thought. By observing the pioneers in our midst, we may come to know the lenses through which they see the world of education and assessment, and thus take the first step toward a new order for education. Some of education’s pathfinders today include the following:

- Indianapolis teachers who are creating an elementary school based on Howard Gardner’s (1989) theory of multiple intelligences.
- Miami, Florida school policy makers who are situating schools at business sites throughout the community.
- Teachers, school support employees, administrators, students, and parents in Rochester, New York, Scottsdale, Arizona, Hammond, Indiana, Los Angeles, California, and growing numbers of other communities around the country, who are learning how to work together to make decisions about education.
- Educators at Prospect School in Vermont who are melding instruction and assessment with student portfolios and professional dialogue.
• Ted Sizer and his Brown University colleagues, consultant Grant Wiggins, and teachers, administrators, and students throughout the nation who are recreating high school education—its curriculum, instructional technologies, and assessments.
• Participants in NEA’s 26 Mastery in Learning Project sites who are framing their own questions about teaching and learning and collaboratively creating their own answers through teacher-directed research.
• Johnson and Johnson at the University of Minnesota who are demonstrating the benefits to both students and teachers of working and learning cooperatively.
• St. Paul high schools that house health clinics.
• Harvard University’s Project Zero staff, Educational Testing Service representatives, and Pittsburgh educators who are devising assessment systems that reflect students’ growth in artistic achievement.

Whether each of these pioneers is aware of the Creative World View or not, each is acting in harmony with it. They are pulling themselves to their futures with their bold visions. They are pursuing their visions in relationship with others. They are in both a state of being and of becoming as they creatively change. Corollary beliefs of these paradigm makers seem to be:

• Learning is meaning-making, pattern discernment, creating.
• Organizations are vehicles for personal and collective empowerment.
• Intelligence is dynamic, multifaceted, and biased towards growth.
• Mind and body are one.

HYPOTHETICAL CREATIVE WORLD VIEW EDUCATION

I invite you to imagine the kind of educational enterprise we could create and the role that assessment would play in it if we, like our contemporary trail blazers, changed the lenses through which we view the world. Try on the lens that reveals all people in the educational organization mutually contributing to the growth of one another. Now, add the one that dissolves the boundary of the schoolhouse. Next, look through the lens that expands your range of vision to include participants of all ages and walks of life. And,
finally, slip into place the lens that discloses the magnificent creative energy of all those people.

With such lenses we might begin to see teachers, students, parents, and all of humanity as capable—regardless of their race or their economic standing or their age. We would probably see “schools” organized as true communities of mutual learning. We could perceive learning as involving real problems, intrinsically rewarding, empowering. And, we would see teaching transformed, synonymous with learning.

Although no new paradigms have securely replaced our current “flat earth” ones—Horace Mann’s teacher as lecturer and controller of information and behavior, classroom as desks and texts confined within four walls, school as one social service agency among many serving a specific geographic area; Eliot’s curriculum as a set of courses in which students serve time; Conant’s school system as consolidated bureaucracy; or Binet’s test—the pathfinders who have been mentioned provide us clues to the ones that may eventually hold sway. Descriptions of the emergent educational paradigms suffer from the same vocabulary problem cited by Kuhn about the clashes between new and old scientific paradigms: Proponents of each view use many of the same words—teacher, classroom, school, curriculum, administration, tests—but with conflicting definitions. Thus, their attempts at discussion of such incommensurable views appear to Clancy (1989) as “conversations of the deaf” (p. 201). Risking this, I offer new definitions inherent in the emerging paradigms.

Teacher: facilitator of learning, guide to potential learning resources, mentor, researcher, collaborative decision maker, coordinator of fellow educators who are of diverse ages—children through retirees—and backgrounds, diagnostician of thinking modes and patterns of growth, student of learning and of general systems, specialist in at least one method of disciplined inquiry whose unique contributions are designed to create synergy with fellow specialists.

Classroom: any physical location one chooses in which to consciously pursue learning, which is equipped with the human resources and technologies that are appropriate to the desired learning.

School: a community of learners, including teachers as defined above, who choose to come together for mutual growth and to serve as the fulcrum for human resource development services to their members and who organize their activities
around an explicitly shared world view whose assumptions are open to continual review.

Curriculum: a set of performance-based, holistic learning outcomes that is aligned with the school's world view and with an expanding knowledge base of human growth and development and which is accompanied by a regularly updated data base of sample learning resources and technologies that are known to contribute to the desired outcomes.

School System: an organization of schools, as defined above, whose decision makers choose to join together for the explicitly defined synergistic effects within their constituent communities of doing so.

And, in such learning communities, what would we see of assessment?

The purpose of teachers' assessment of students in schools aligned with this creative world view would be empowerment of both students and teachers. It would be designed to reveal students' methods for making sense of the world—their patterning styles; their "intelligences," to borrow Howard Gardner's term; their conceptions and misconceptions of various operations; and the degree to which they convey their integrated understandings through a variety of performances.

The role of teachers in fulfilling the purpose of such assessment would be as full partners in determining a shared vision for their school that includes holistic performance goals for students and plans for how to foster and evaluate student performance. Their role would also include full partnership with other educational specialists—researchers, test developers, for instance—and with their fellow learning community members—students, colleagues, parents, citizens—in the development and administration of assessment processes that were consistent with this world view and in the application of their findings from these assessments.

The responsibility of teachers for fulfilling such a purpose? To assume the authority of a full partner and accept the responsibility for exercising such authority.

PREPARATION OF TEACHERS

What kind of education of teachers is implied by the picture of education and assessment revealed through new world-view lenses?
Because my work entails the development of practicing teachers, I will focus on the concurrent courses of action that this tumultuous between-paradigms phase has engendered within the Arizona Education Association. We are engaged in the following efforts as we “convert” from one set of paradigms to another:

1. Encouraging bold new visions of education by creating such visions for our own organization through strategic planning and by providing resources to members that encourage them to examine their own world views of and assumptions about education. For instance, we’ve just produced a booklet on restructuring education in Arizona and are working with local association leaders on ways to make use of it in their districts and with their communities to commence the dialogue that must precede the development of new, shared visions of education. We are organizing viewings and discussions of Joel Barker’s video on paradigms. And, we offer grants (Learning Improvement through Faculty Teams--LIFT--grants) to members who are undertaking leading edge transformations of teaching and learning.

2. Offering training in site-based decision making to members and to teacher-administrator audiences to assist them in functioning effectively as interactive work teams. The National Education Association has produced in-depth reference and training materials on this topic in consultation with experts in participatory management. Using these materials as a springboard, we also are creating experiential learning opportunities for people to hone their skills as facilitators of consensus decision making.

3. Forming alliances (or organizing at a higher level of complexity) with other groups to transform education. AEA has worked cooperatively with the state legislature to adopt broad goals for education in Arizona and with the state department of education to support and publicize a comprehensive new accountability system of multiple indicators to monitor progress towards those goals, that includes the assessment of student performance of complex problem-solving tasks. This, in turn, is leading to alliance-formation with subject-matter and professional education organizations and with higher education institutions to create the staff development opportunities that
teachers will need to align curriculum, instruction, and classroom organization with the new performance outcomes. Because the time required for staff development, collaborative decision making, and community-based dialogues about education depends on financial resources, AEA is also part of a coalition of educators, policy makers, parents, business leaders, and citizens that has secured 200,000 signatures to place a school funding initiative on Arizona's November 1990 ballot.

Perhaps these actions in Arizona, and similar ones around the country, will serve as an impetus to colleges and universities, to state education agencies, and to accreditation bodies, to revamp their own approaches to the education of teachers. Something will. Something must.

HOW TO BEGIN

John Goodlad reminds us that the future does not arrive full blown, but rather is defined by the small decisions we make each day. The new sciences of quantum mechanics and cognition reveal that every one of us is in the process of creating reality, that, in fact, "the possibilities we imagine for anything actually make up half of its reality" (Jarman & Land, 1992, p. 52) and the act of cognition does not simply mirror an objective reality "out there," but instead is an active process, rooted in our biological structure, by which we actually create our world of experience (Maturana & Varela, 1988). Thus, the most important action we each can take is to redefine our present idea of reality by exchanging our restrictive world-view lenses for those that expand and clarify our field of vision.

Redefining Our Present Idea of Reality. First, we must know what our present idea of reality is—what the power of our current world-view lenses is—and then how it defines our "edge of the world." That entails identifying our own beliefs, assumptions, and attitudes about how the world works and, thus, how education works. Three of the ways in which we can do so are the following:

- Be scrupulously honest with ourselves.
- Ask other people what blind spots they perceive in us. (We can't see our own blind spots because we can't see what we can't see.)
- Analyze what it is we do see because our perceptions tell us what we believe. For instance, if I see teachers and
administrators “fighting over who’s in charge,” it reveals to me a belief that we are separate from each other, disconnected. If, however, I see teachers and administrators exhibiting that same behavior as “forming new relationships of greater complexity,” it reveals my beliefs that all things and everyone is connected—at all times and instantaneously—and that disorder precedes the formation of connections at deeper, broader, more complex levels.

Next, we need to challenge, expand, and deepen our world view. Ways to do that include the following:

- reading, analyzing and comparing the familiar and the unfamiliar—professional literature and Greek philosophers, educational research and chaos theory, textbooks and electronic data bases;
- talking rigorously about familiar and unfamiliar ideas with people who are familiar and unfamiliar to us—a colleague and a nuclear physicist, a parent and a cognitive scientist, a student and a musical composer, a family member and a top-performing athlete; and
- conducting our own comparative research on the familiar and the unfamiliar.

Third, we have the responsibility of helping all education stakeholders to expand their world view, as well. It is only when we have all exchanged our Ben Franklin spectacles for lenses that allow us to see much broader horizons, that we will have completed the action of redefining our sense of reality and be capable of creating a new reality for education.

*Creating A New Educational Reality.* Based on a world view that is more closely aligned with what is known about the current scientific facts of life and about perception, cognition, and human development, we then must ask ourselves and fellow stakeholders two questions: “What purpose do we want education to serve in this community?” and “What do we want students to know and be able to do as a result of participation in this education process?” This constitutes the vision creation process—President Bush’s “vision thing.”

Next, as Stanford Professor Larry Cuban (1989) suggests, we must ask, “What should we do to help students reach these ends?” Answering this query will involve looking at how to organize the enterprise of learning, how to structure curriculum, how to employ instruction, how to use time as a resource, and how to assess and build on students’ learning strengths.
Only then are we ready to ask the accountability-related question, “How will we know that we’re fulfilling our school’s purpose and that students are reaching the desired ends?” The kind of accountability system developed in response to this will be one that allows the educational enterprise—the “Ship of Schools”—to self-correct the course of its voyage because finally it knows where it is, where it is going, and is powering itself to its future, its new reality.

Who Should Begin? Any one person can begin. As the scientists studying the new field of chaos tell us, a butterfly flapping its wings in China theoretically can affect the wind patterns in New York City. And, as Jarman and Land (1992) convey the findings of research on major social changes, “Five per cent of a population needs to change before the established leaders begin to take notice that something new is happening. Once that intrepid 5% convinces another 15%, then a rapid and unstoppable momentum shifts the other 80%” (p. 68).

I invite you to return to the quiz about your beliefs with which I began this presentation and to keep in mind this ancient Sufi parable reported by Maturana and Varela (1988) in their exploration of the biological roots of human understanding:

A story is told of an island somewhere and its inhabitants. The people longed to move to another land where they could have a healthier and better life. The problem was that the practical arts of swimming and sailing had never been developed—or may have been lost long before. For that reason, there were some people who simply refused to think of alternatives to life on the island, whereas others intended to seek a solution to their problems locally, without any thought of crossing the waters. From time to time, some islanders reinvented the arts of swimming and sailing. Also from time to time a student would come up to them, and the following exchange would take place:

“I want to swim to another land.”
“For that you have to learn how to swim. Are you ready to learn?”
“Yes, but I want to take with me my ton of cabbages.”
“What cabbages?”
“The food I’ll need on the other side or wherever it is.”
“But what if there’s food on the other side?”
“I don’t know what you mean. I’m not sure. I have to bring my cabbages with me.”
“But you won’t be able to swim with a ton of cabbages. It’s too much weight.”
“Then I can’t learn how to swim. You call my cabbages weight. I call them my basic food.”
5. TEACHERS’ ASSESSMENT OF STUDENTS

“Suppose this were an allegory and, instead of talking about cabbages we talked about fixed ideas, presuppositions, or certainties?”

“Hummm... I’m going to bring my cabbages to someone who understands my needs.” (pp. 249-250)

Are we willing to let go of the ideas that weigh us down and that blind us? Believing is seeing. Act as though you believe one person can make a difference. Act as though you can create a new reality for education. And watch reality shift. See it happen before your eyes.

REFERENCES


Cuban, L. (1989, October). First and second order change in education. Presentation made at the annual National Education Association meeting of state affiliate directors of instruction and professional development, San Diego, CA.


Teachers' testing practices, as reflected in such activities as stating desired learner outcomes, grouping pupils, instigating study activities, and providing feedback for monitoring teaching and learning, are an integral component of models of instruction (Brophy & Good, 1986; Rosenshine, 1985). The testing and assessment process within learning models is variously described as providing practice, review, consolidation of learning, knowledge of results, feedback for redirecting efforts, feelings of accomplishment, a focus for efforts, etc. Relatedly, Crooks (1988) asserts that testing/evaluation is one of the most potent forces influencing education. Also, Elton and Laurillard (1979), in describing the impact of classroom testing upon pupils, stated that the surest way to change pupil learning behavior is to change pupil assessment.

Contrary to the common perception that testing plays an essential role in the teaching and learning process, actual elements of the evaluation schemas that teachers institute have received less research attention than most other aspects of education (Crooks, 1988). Further, the research of testing has been focused primarily upon standardized testing rather than upon the much more prevalent teacher-devised testing, and those studies that have addressed teacher-made tests and teachers' testing practices have predominantly used teacher self-report data-gathering procedures. As a consequence, these limited and narrow research efforts have resulted in testing professionals
knowing little about the nature and quality of teacher-made tests, about how these tests are used within the classroom teaching-learning process, and about the adequacy of teachers’ testing knowledge and skills (Stiggins, Conklin, & Bridgeford, 1986).

The purpose of this chapter is to provide a review of the professional literature devoted to testing in the classroom in order to ascertain what testing knowledge and skills K-12 classroom teachers ought to have; what testing practices ought to be used to facilitate classroom learning; what is known about teachers’ actual testing knowledge, skills, and practices; and what implications for the measurement profession are suggested by any discrepancies identified between teachers’ desired and actual testing knowledge, skills, and practices. More specifically, this chapter is focused upon teachers’ testing knowledge, practices, and skills, and is organized around the following five questions:

1. What should the nature and extent of K-12 classroom teachers’ testing knowledge, skills, and practices be, as indicated by the findings from research on testing in the classroom and by the expectations and advice of the professional measurement and educator communities?

2. What is the nature and extent of the school community’s support for testing in the classroom? What are the school community’s perceptions regarding the adequacy of teachers’ testing knowledge and the adequacy of teachers’ training in testing? And to what extent are resources such as duplication services available in schools to assist teachers in meeting their testing responsibilities?

3. What is the extent of K-12 classroom teachers’ testing knowledge as revealed through their reported testing practices, beliefs, and attitudes?

4. What is the extent of K-12 classroom teachers’ testing knowledge and skills as revealed through paper-and-pencil assessments; through proficiency ratings of teachers’ testing competencies, completed by the teachers themselves and by principals and supervisors; and through direct assessments of teachers’ test construction skills as revealed on their formal teacher-made tests?

5. And finally, how do K-12 classroom teachers’ testing knowledge, skills, and practices measure up, and what recommendations for the measurement profession are suggested by the findings from the review of the research literature pertaining to testing in classroom settings?
6. TEACHERS' TESTING PRACTICES

DESIRED TESTING KNOWLEDGE, SKILLS, AND PRACTICES

Research on Classroom Testing

The intent of the writers in this section, and throughout most of the chapter, is not to describe the measurement research literature in detail but to describe briefly the research findings with implications for K-12 classroom teachers' testing practices. The reader should note that other chapters in this book provide more complete discussions of several of the topics presented in this chapter and that a few rather extensive literature reviews of these or closely related topics also exist (e.g., the reviews provided by Balch, 1964; Bangert-Downs, Kulik, & Kulik, 1988; Crooks, 1988; and Kulik & Kulik, 1988).

Research of various variables associated with classroom test characteristics and classroom testing practices has been conducted throughout much of the 20th century. Even though this research has been conducted predominantly in college classrooms, Crooks (1988) has argued that the findings from these studies have been sufficiently replicated in K-12 classrooms to warrant generalization to these latter classrooms, with a few cautions. For example, he noted that some inconsistencies in findings are not uncommon in this research literature and that some testing conditions, such as testing frequency, appear to have a greater positive impact upon younger and less able pupils.

Tests guide and instigate effort. It is rather clear from the research on the impact of testing upon students' learning, often involving interviews of pupils, that pupil study is instigated by an announced test and is focused primarily upon content that they anticipate will appear in the test. In regard to this impact of tests upon pupils, Rogers (1969) stated that classroom tests inform learners of the real aims of a class, at least so pupils believe.

The directing of pupil study efforts toward content that is tested may have desirable or detrimental effects upon learning, depending upon how well the test directs pupils to desired outcomes. In order for tests to properly direct pupil study efforts, the testing community advises teachers to use test specification tables to better link test questions to desired learner outcomes. This matching of test items with desired outcomes frequently is not done, and the resulting absence of match between content of classroom tests and more significant course content is often recognized by both teachers and pupils. For example, Snyder (1971) reported that students' primary goal in planning their study efforts was performing well on course examina-
tions, although they commonly saw this goal as conflicting with true learning of the subject matter. Snyder referred to this adverse impact of poorly designed tests upon pupil learning efforts as the hidden curriculum in education.

**Question type influence.** A number of researchers have reported that pupils vary their pattern of study when informed of the types of test questions to appear in a scheduled classroom examination. Balch (1964), after a review of teacher-instigated testing studies, concluded that pupils' awareness of the nature of the classroom test to be administered and the provision of feedback regarding pupils' performance following a test are the two most potent testing variables influencing classroom learning. He described pupil study strategies as focusing on details when preparing for objective tests, and as searching for relationships and main points when preparing for essay tests. In more recent research, D'Ydewalle, Swerts, and DeCorte (1983), Gay (1980), and Sax and Collett (1968) have reported similar findings. In response to this research, testing specialists commonly advise teachers to use a variety of question types on their classroom tests, when appropriate for the content to be examined, to encourage pupils to use more varied study patterns.

**Testing frequency.** Bangert-Downs, Kulik, and Kulik (1988), after reviewing a number of studies of classroom testing frequency, concluded that pupils in classes with no tests scheduled were clearly disadvantaged, that moderately frequent tests appear to best facilitate pupil achievement, and that as test frequency in a course increases pupil achievement benefits resulting from these additional scheduled tests begin to diminish. They also noted that the facilitating effect of frequent testing upon pupil achievement appears to be consistent across subject content fields, to be more beneficial for less able pupils than for more able pupils, and to be more beneficial under certain testing conditions, such as the provision of feedback related to pupil performance on tests following the examination period. Testing also has been found to be superior to equal amounts of classroom time spent on content-reviewing activities in facilitating pupil achievement, and pupils report that they prefer and learn more when relatively frequent tests are scheduled during a course (Guza & McLaughlin, 1987; Halpin & Halpin, 1982; Marso, 1970a; Monk & Stallings, 1971; Nungester & Duchastel, 1982; Peckham & Roe, 1977).

**Test administration mechanics.** Research suggests that announced and carefully administered and monitored classroom tests, for which content and format are described to pupils prior to administration,
typically produce higher pupil performance, less pupil cheating, and reduced pupil test anxiety (Bushway & Nash, 1977; Carrier & Titus, 1981; Hill & Wigfield, 1984; Saigh, 1984; Szafran, 1981; Trentham, 1975). Conversely, unannounced tests, carelessly administrated tests, poorly monitored tests, and tests perceived by pupils to be unfair not only adversely impact upon student performance but tend to heighten test anxiety and encourage cheating.

Test feedback. The prompt return of classroom tests with the provision of knowledge of results or other forms of pupil feedback, such as discussion of questions missed, tends to increase pupil achievement (Kulik & Kulik, 1988; Marso, 1970a; Wexley & Thornton, 1972). This positive relationship between the provision of test feedback and pupil achievement appears to hold at all pupil grade and ability levels. Research also suggests that the return of scored exams in the class period following the exam should be construed as prompt feedback, for the presentation of knowledge of performance immediately following pupil responses to individual test questions can be distracting to the extent that pupil achievement is impaired (Bridgeman, 1974).

Question difficulty and arrangement. Research of the impact of test question difficulty and of test question arrangement upon pupil achievement has been less conclusive than the findings from the research of many other aspects of testing. The authors of preservice educational measurement textbooks persist in recommending that questions be arranged from easy to difficult on teacher-made tests, even though neither research findings not motivational principles provide clear support for this advice. Similarly, teachers are commonly advised when constructing formal teacher-made tests that test difficulty should be approximately 50%, after adjustments for probability of guessing relative to question types used, in order to assure an acceptable level of test reliability (Gronlund & Linn, 1990; Mehrens & Lehmann, 1984).

Motivational principles and logic suggest, however, that pupils' study efforts would be more effectively rewarded by a moderately high level of pupil success on teacher-made tests. Pupils having experienced one or more very difficult tests in a course are less likely to be motivated to persist in their course study efforts if they assume that all subsequent tests in the course will be as difficult or more difficult than if they assume that some subsequent tests in the course will be sufficiently less difficult to allow them to experience more success. Similarly, students having experienced four or five consecu-
tive and very difficult questions on a test are less likely to be motivated to persist in their testing efforts if they assume that all subsequent questions on the test will be increasingly more difficult than if they assume that some subsequent questions on the test will be less difficult.

The research of test question difficulty arrangements, such as random placement or easy-to-difficult placement within tests, indicates that arrangement patterns generally have little impact upon student test performance on teacher-constructed tests (Klimko, 1984; Marso, 1970b; Monk & Stallings, 1970; Newman, Kundert, Lane, & Bull, 1988). On the other hand, limited research suggests that the level of total test difficulty may influence pupils’ test preparation efforts and achievement. This latter research suggests that moderately difficult (as compared to more difficult) teacher-made tests increase pupil study efforts and achievement. Thus, motivational principles and limited research suggest that K-12 classroom teachers ought to be advised when preparing formal tests to construct moderately as opposed to more difficult (e.g., 70% item difficulty average rather than 50%) tests and to arrange questions in random difficulty order within question type groupings.

**Test cognitive demands.** In the introduction to the December 1989 issue of *Educational Researcher*, which was devoted to educational assessment and the enhancement of pupil higher order thinking skills, Nickerson (1989) pointed out that the conflict between “studying for the exam” and “learning for learning’s sake” dissipates when test questions are closely related to desired learning outcomes and also are functioning within a desirable range of cognitive levels. A common criticism of teacher-made tests, however, is that they tend to function almost exclusively at the recall or knowledge cognitive level (Fleming & Chambers, 1983; Marso & Pigge, 1988a), and studies of K-12 classroom teachers’ testing practices indicate that teachers generally do not use test specification tables to better match test questions with content objectives (Gullickson & Ellwein, 1985; Marso & Pigge, 1988a).

There appears to be a consensus among measurement specialists that teacher-made tests need to function at higher cognitive levels to assure attainment of instructional goals and to promote higher level pupil thought processes. Similarly, teachers, principals, and supervisors also report that they believe it is important for teacher-devised tests to function at higher cognitive levels (Marso & Pigge, 1987a). Despite this apparent consensus among these various professionals, not only does research suggest that teachers’ tests do not function at
higher cognitive levels, but there appears to be no empirical evidence linking the cognitive functioning level of teacher-made tests to pupil achievement or to pupil thought processes.

Measurement Profession Expectations of Teachers’ Testing Knowledge

During the late 1980s, the measurement profession, through the efforts of the National Council of Measurement in Education, the American Association of Colleges for Teacher Education, the American Federation of Teachers, and the National Education Association, developed standards for classroom teachers’ competence in pupil assessment that were published in 1990. As these standards are described in detail in another chapter, they will be dealt with very briefly here.

For this chapter, the significance of the professional standards for teachers’ competence in student assessment is this: The standards represent the measurement profession’s perceptions of what classroom teachers ought to know about testing. The measurement profession’s standards for teacher competence in the assessment of pupils indicate that classroom teachers ought to be knowledgeable about and proficient in:

- the selection of appropriate assessment methods for making various instructional decisions
- the development of assessment devices or procedures appropriate for making various instructional decisions
- the appropriate administration and scoring of assessment devices and the appropriate interpretation of the results of classroom assessments
- the appropriate use of classroom assessment results in making instructional and related decisions about pupils and school curricula
- the appropriate communication of classroom assessment results to pupils and related audiences
- the identification and appropriate response to ethical and legal issues and concerns related to classroom assessments, such as honoring pupil and family privacy rights and privileges, avoiding discriminatory practices, and alleviating potential negative labeling effects

Educators’ Expectations of Teachers’ Testing Knowledge

Teachers report that they place more reliance on informal than formal assessments in making K-12 classroom decisions (Gullickson,
1985; Linn, 1990; Salmon-Cox, 1981). Teachers also report a need for test construction skills and a need for formative (in contrast to summative) evaluation, but they report little need for measurement statistics and for knowledge of legal issues associated with testing in K-12 classrooms (Gullickson, 1986a). Teachers further perceive teacher-made tests and informal observations of pupils to be useful in making day-to-day instructional decisions, but they consider previous teaching experiences to be more useful than test scores in planning instruction for the school year (Dorr-Bremme, 1983).

Borg, Worthen, and Valcarce (1986) and Marso and Pigge (1987a) found the K-12 classroom teachers rated more highly their need for measurement skills closely associated with instruction than their need for skills such as writing structurally sound test questions. Similarly, Newman and Stallings (1982) found that teachers reported heavy reliance upon their self-constructed tests for making decisions about activities most closely related to instruction, such as diagnosing pupil strengths and weaknesses, assessing pupil progress, and assessing pupil mastery of units of instruction; whereas the teachers reported somewhat less reliance upon teacher-constructed tests for assigning grades.

The data presented in Table 1 are illustrative of classroom teachers', building principals', and supervisors' ratings of classroom teachers' need for a variety of testing competencies (Marso & Pigge, 1987a). As did the teachers in previously noted studies, these classroom teachers reported relatively little need for measurement statistics. The teachers reported a high need for competencies involving instructional use of test results (grading and scoring activities, reteaching, identifying pupil strengths and weaknesses) and test validity-related competencies (matching questions with objectives, writing questions that measure higher thinking, making tests that reflect what was taught, and measuring true progress of pupils).

Rather surprisingly, the teachers reported a rather low need for question-writing skills that could be deemed necessary to attain the test validity and instructional uses they rated highly. Similarly, the teachers rated rather low the need for competency in selecting good test questions from sources such as teacher manuals. Collectively, these teachers' ratings of needed testing competencies suggest relatively little teacher concern for question structural quality as compared to other question validity concerns, and direct analyses of these teachers' self-constructed tests revealed frequent violations of common question writing guidelines. These violations, in part, may have
### Table 1. Means and Ranks of Teachers' and Administrators' Ratings of Classroom Teachers' Need for Selected Testing Competencies

<table>
<thead>
<tr>
<th>Testing Competencies or Skills</th>
<th>Classroom Teachers (N=313)</th>
<th>Principals and Supervisors (N=580)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Mean</strong></td>
<td><strong>Rank</strong></td>
</tr>
<tr>
<td>1. Writing good multiple-choice questions</td>
<td>3.33</td>
<td>20</td>
</tr>
<tr>
<td>2. Writing good completion questions</td>
<td>3.53</td>
<td>19</td>
</tr>
<tr>
<td>3. Writing good matching questions</td>
<td>3.54</td>
<td>17.5</td>
</tr>
<tr>
<td>4. Writing good true-false questions</td>
<td>3.31</td>
<td>21.5</td>
</tr>
<tr>
<td>5. Writing good essay questions</td>
<td>3.20</td>
<td>24</td>
</tr>
<tr>
<td>6. Scoring essay questions</td>
<td>3.24</td>
<td>23</td>
</tr>
<tr>
<td>7. Identifying good and poor questions for future tests</td>
<td>4.03</td>
<td>9.5</td>
</tr>
<tr>
<td>8. Writing questions in harmony with school and class goals</td>
<td>4.01</td>
<td>11</td>
</tr>
<tr>
<td>9. Stating objectives sufficiently clear to suggest test items</td>
<td>3.88</td>
<td>15</td>
</tr>
<tr>
<td>10. Writing test questions that demand higher thinking processes</td>
<td>3.81</td>
<td>16</td>
</tr>
<tr>
<td>11. Constructing tests that represent true student progress</td>
<td>4.18</td>
<td>7</td>
</tr>
<tr>
<td>12. Use of less formal assessments: checklists, ratings, etc.</td>
<td>3.31</td>
<td>21.5</td>
</tr>
<tr>
<td>13. Use of observations (visual) to assess and guide learning</td>
<td>4.03</td>
<td>9.5</td>
</tr>
<tr>
<td>14. Use of sociometric, guess who, and related techniques</td>
<td>2.71</td>
<td>25</td>
</tr>
<tr>
<td>15. Selecting good test questions from teacher manuals</td>
<td>3.54</td>
<td>17.5</td>
</tr>
<tr>
<td>16. Setting up readable, scorable, and attractive tests</td>
<td>3.94</td>
<td>14</td>
</tr>
<tr>
<td>17. Making tests reflect what is covered in text and class</td>
<td>4.35</td>
<td>2</td>
</tr>
<tr>
<td>18. Calculation of means, standard deviations, reliability, etc.</td>
<td>2.49</td>
<td>26</td>
</tr>
<tr>
<td>19. Interpreting test scores and student progress</td>
<td>4.00</td>
<td>12</td>
</tr>
</tbody>
</table>

(Continued...)

TABLE 1. (continued)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20. Identifying individual and class strengths and weaknesses</td>
<td>4.25</td>
<td>4</td>
<td>4.41</td>
<td>5</td>
</tr>
<tr>
<td>21. Determining what needs to be retaught after tests</td>
<td>4.20</td>
<td>6</td>
<td>4.53</td>
<td>1</td>
</tr>
<tr>
<td>22. Use of tests and grades to positively influence learning</td>
<td>3.99</td>
<td>13</td>
<td>4.30</td>
<td>11</td>
</tr>
<tr>
<td>23. Calculating end of term grades from term work</td>
<td>4.29</td>
<td>3</td>
<td>3.95</td>
<td>18</td>
</tr>
<tr>
<td>24. Grading tests, papers, projects, homework, etc.</td>
<td>4.44</td>
<td>1</td>
<td>4.09</td>
<td>15</td>
</tr>
<tr>
<td>25. Deciding importance of tests, papers, etc. in grading</td>
<td>4.23</td>
<td>5</td>
<td>4.18</td>
<td>14</td>
</tr>
<tr>
<td>26. Deriving information from tests to guide students</td>
<td>4.04</td>
<td>8</td>
<td>4.36</td>
<td>7</td>
</tr>
</tbody>
</table>

*Means were derived from a 5-point Likert scale where 5 = high.

resulted from the teachers’ low regard for test question structural quality (Fleming & Chambers, 1983; Marso & Pigge, 1988a).

Teachers’ perceptions of their relative need for various measurement competencies were found to be very similar to those of the principals and teacher supervisors. These administrators and the teachers differed from one another, however, in their ratings of teachers’ needs for essay testing, classroom observation, and pupil grading-related competencies. The teachers rated their need for competencies related to classroom observations and pupil grading considerably higher than did the principals or supervisors; whereas the administrators perceived more need for teachers’ essay testing skills than did the teachers. The finding of teachers rating more highly their need for those testing competencies they perceived to be needed to meet the day-to-day demands of the classroom than they rated other testing competencies is consistent with the findings from studies noted previously.

The findings from the review of the research literature related to classroom testing practices, and to the educational and measurement professions’ perceptions of testing competencies needed by teachers to function successfully in classrooms, are summarized in Table 2. Considerable research evidence and professional consensus support these statements, although the extent of evidence and consensus varies among the individual statements.
## Table 2. Desired Teachers' Classroom Testing Knowledge, Skills, and Practices Derived from Professional Consensus, Published Standards, and Classroom Testing Research

1. Select appropriate assessment methods for making various instructional decisions.
2. Construct appropriate assessment devices for making various instructional decisions.
3. Appropriately administer and score assessment devices and interpret the results of classroom assessments.
4. Appropriately use classroom assessment results in making instructional and curricula decisions.
5. Appropriately use classroom assessments in making decisions about pupils and in assigning pupil grades.
6. Appropriately communicate assessment results to pupils and related audiences.
7. Identify and appropriately respond to ethical and legal issues and concerns related to assessment.
8. Interpret test scores within the context of other pupil data.
10. Appropriately interpret common scores derived from standardized tests.
11. Arrange questions in random difficulty within similar question type groupings within an attractive and readable test format in preparing teacher-devised tests.
12. Calculate means and standard deviations of test scores and interpret these indices appropriately in communicating test results to pupils and in assessing the quality of teacher-made tests.
13. Construct tests sufficiently difficult to achieve reliability but sufficiently easy to promote learning and study efforts.
14. Use a variety of question types in making classroom tests consistent with the nature of the course content to be measured.
15. Use a test specification table or similar process to assure the use of questions measuring at a variety of cognitive levels and a match of questions with instructional objectives.
16. Select and construct test questions in accord with commonly accepted question construction guidelines.
17. Use basic item analysis procedures to direct reteaching activities and to improve future tests and instruction.
18. Describe, announce, frequently schedule classroom tests, monitor pupils taking tests, and promptly return and discuss with pupils their performance on the tests.
19. Select and construct test questions functioning in a diverse range of cognitive levels.
21. Construct, use, and interpret less formal pupil assessment data gathering procedures such as checklists, product and performance rating scales, scociometric techniques, and anecdotal records.
22. Combine and appropriately weight test scores and the results of other assessments in order to make decisions about pupils and to accurately assign pupil marks.
Measurement Community Resources and Perceptions

Until the standards for teacher competence in the assessment of pupils described in the preceding section were published in 1990, the testing community had not provided clear expectations or standards regarding classroom teachers' testing competence. Conversely, the existence of statements of standards for standardized testing can be traced back to the mid-20th century. These statements are currently conveyed in the 1985 *Standards for Educational and Psychological Testing*, jointly developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA, APA, NCME, 1985). More recently, these standards were supplemented by the 1988 Code of Fair Testing practices in Education, also jointly sponsored by these three professional organizations. The Code was designed to complement the earlier standards and differs from the standards in audience addressed and purpose. It is focused just upon standardized educational testing but addresses the practices of both test developers and test users. Its primary role is to address test and test score misuses that have tended to generate far more public criticism than have questions about test quality itself (Diamond & Fremer, 1989).

Neither the code nor the standards address teacher-devised testing. Frisbie and Friedman (1987) did make an effort to show a relationship between the standards and teacher-devised testing; however, the result of their effort was illustrative rather than enumerative in scope. Thus, it appears that the measurement community has provided less professional guidance for and (as noted previously) less research of teacher-made testing than it has for standardized testing. This relative neglect of teacher-devised testing has occurred in spite of the fact that the measurement profession perceives teacher-made tests, not standardized tests, to be the dominant influence in K-12 classrooms (Stiggins, 1985).

Even though the measurement community appears to have provided less research support and professional guidance for teacher-devised testing in contrast to standardized testing, it appears to have considerable doubts about the testing knowledge, skills, and practices of educators. For example, Diamond and Fremer (1989) noted that the Institute for Research on Teaching, which coordinated the development of the previously described fair testing code, was particularly
critical of the inadequate training of educational personnel in the interpretation and use of tests.

Further, the questioning of the adequacy of teachers’ testing knowledge is not a recent occurrence. Gullickson (1986b) traced the professional concern about the adequacy of teachers’ testing and evaluation knowledge back to Conant’s book, *The Education of American Teachers* (1963); to Mayo’s survey of principals, superintendents, and professors about what teachers ought to know about testing (1964); and to Mayo’s testing of teacher candidates about what they did know about classroom testing (1967). The measurement community’s questioning of the extent of teachers’ testing knowledge is also widespread, as Gullickson cited several recent studies revealing the inadequacy of teachers’ testing skills and knowledge. Wanous and Mehrens (1981), in describing a strategy for helping teachers develop testing knowledge, also commented about the inadequacy of both teachers’ testing knowledge and training. In addition, Rudman, Kelly, Wanous, Mehrens, Clark, and Porter (1980), following an extensive review of research on testing in classrooms, concluded that many have doubts about the adequacy of teachers’ testing knowledge.

**School Community Resources and Perceptions**

The extent of the availability of testing expertise, and of other forms of support for teacher-devised testing in the schools, appears to be as bleak as the measurement community’s perceptions of the adequacy of teachers’ testing competencies. Ruddell (1985), after conducting interviews of school principals, school district central office staff, state legislators, and classroom teachers, concluded that they all possessed very limited knowledge about tests and test score interpretation concepts, such as the standard error of measurement.

Marso and Pigge (1990) conducted a survey of school-district-designated directors of standardized testing and found that many school testing directors themselves have limited training in testing and evaluation. Contrary to the expectations stated in the *Standards for Educational and Psychological Testing*, many of the testing directors, when queried about support services they provided for classroom teachers, reported that they were not responsible for encouraging the use of standardized test results in their schools, for training teachers to proctor standardized tests, and for training teachers to better interpret scores from standardized tests.

Marso and Pigge also found that many of the testing directors reported increased demands on their time, resulting from added
responsibilities for the management of mandated statewide pupil competency testing, thus undoubtedly also reducing the testing directors’ opportunities for providing teachers with testing expertise or support services. These researchers concluded it is probably safe to assume that if testing directors do not provide basic testing support services for teachers, at least in the smaller school districts, these essential services are probably not being provided in the schools. They reached this conclusion partly on the assumption that no one else in these schools would likely have this responsibility or the expertise to deliver such services.

Relatedly, Stiggins (1985) noted that few school administrators have the training or the experience necessary to help teachers with classroom testing or related responsibilities. Further, Marso and Pigge (1989c) reported negative correlations between principals’ and supervisors’ ratings of teachers’ various question-type writing skills (e.g., ability to write multiple-choice and other types of questions) and the observed levels of the adequacy of teachers’ various question-writing skills as displayed on their self-constructed tests. As the adequacy of the teachers’ test question-writing skills in this study was judged on the basis of the frequency that common test construction guidelines were violated, this finding may suggest that school administrators, who themselves tend to have little or no training in testing, may not be able to identify violations of test question-writing guidelines when examining teacher-constructed tests.

Lambert (1980-81) collected opinions about teachers’ attitudes, training, and knowledge about teacher-made and standardized tests from a national sample of state legislators, state teacher association officials, and deans of colleges of education. He found both agreement and divergence between and within these three samples. For example, approximately one third of the deans reported that their colleges did not offer a measurement course for their teacher candidates and that they had no intention of doing so. Nevertheless, all three groups agreed with one another that classroom teachers have a negative attitude toward standardized tests, that teachers should know more about tests, and that it is very important for teachers to construct superior tests for the assessment of their pupils. Lambert concluded that all three groups needed to know more about the value and limitations of tests.

Relatedly, Sproull and Zubrow (1981) found that central administrators of schools do not perceive the management of standardized testing as being a very important administrative function and that few schools have formal testing offices as such to manage these activities;
Tyler and Sheldon (1979) reported a relatively unclear and weak linkage between standardized tests and teachers' use of the results from these tests in their instruction; Marso and Pigge (1989b) found that principals and teacher supervisors believe standardized testing skills are less needed by teachers than testing skills associated with teacher-devised tests or pupil competency tests; and Kinney, Brickell, and Lynn (1988) found that building principals commonly do not perceive the need for testing and measurement specialists to be involved in the selection of standardized tests or in the construction of locally developed tests designed for district-wide use.

In regard to the extent of direct support available for teachers' testing activities, Marso and Pigge (1988d) asked over 800 teachers, principals, and supervisors to report on the availability of selected school resources to support classroom teachers' testing responsibilities. They found that basic typing and duplication services were not consistently available in 50% of the schools, grade assignment guidelines were not available in 50% of the schools, and basic computer services (e.g., test scoring, item pools, item analyses, etc.) were not available in approximately 75% of the schools.

Dorr-Bremme (1983), using questionnaire and interview procedures to gather data from a national sample of school staff in 114 school districts, reported that most teachers do not receive in-service training or assistance of other types in selecting, developing, and using tests. Rather significantly, these researchers found a positive relationship between teachers' attitude toward school testing and the amount of school support for testing in the form of expressed principal interest, resources available for testing, and availability of in-service teacher training related to testing. In school districts where these testing support services were extensive, teachers' attitude toward testing was positive; in school districts where these resources and services were very limited, teachers' attitude toward testing was less positive.

In other studies related to the availability of support for testing, Gullickson (1984) found that teachers reported having little assistance in the form of aides or professional staff in the preparation, analysis, scoring, or interpretation of teacher-made tests. And in another study providing evidence of schools' poor communication about the purpose of (if not the poor management and support of) testing, Salmon-Cox (1981) reported that neither school administrators nor teachers perceived that they were the group primarily benefiting from standardized testing. Teachers perceived standardized testing as prima-
rily benefiting administrators, whereas administrators perceived benefits from testing primarily accruing to the instructional staff.

School Community Support of Training for Testing

Hermanowicz (1980) argued that a major component in teacher preservice education ought to be training in the development of measurement and evaluation proficiencies. Practicing teachers themselves report that assessment of pupils is a key element in the instructional process, and measurement specialists such as Stiggins, Conklin, and Bridgeford (1986) and Dorr-Bremme (1983) have provided information describing how classroom teachers do integrate testing within their day-to-day instructional practices. Further, Schafer and Lissitz (1987) reported an increasing awareness of the importance of teachers’ pupil assessment skills within the educational community, as evidenced by the positive positions taken by the two major national teacher organizations on pupil assessments and by the inclusion of testing as one of the five skill components measured by the recently revised National Teachers Examination.

Despite the educational community’s increasing awareness of teachers’ need for pupil assessment competencies in providing instruction, considerable evidence exists that a significant proportion of professional school personnel receive little or no formal training in measurement and evaluation. After conducting a survey of 438 institutions of higher education, Schafer and Lissitz (1987) found that only approximately one third of the educational personnel preparation programs required a measurement course for certification. Even more disconcerting, they found that just approximately 25% of the elementary and secondary teacher preparation programs required a measurement course. They further noted that, although administrators are expected to serve as instructional leaders in schools, the administrator education programs were least likely of all preparation programs to require measurement training. Among the advanced certification programs for educators, they found that only the counseling programs are very likely to have a measurement course requirement.

Gullickson and Hopkins (1987) conducted a regional survey of 99 colleges of education and found that approximately one half of the colleges provided a separate measurement course for their preservice teachers, whereas the other colleges provided measurement instruction as a unit within another course. Roeder (1973), following a survey of 860 colleges of education conducted some years ago, reported that somewhat fewer than one half of the training programs
required a separate tests and measurement course for their elementary education candidates.

Relatedly, Green and Williams (1989) found that teachers with more training in measurement reported scheduling teacher-made tests more frequently in their classrooms and using the results of standardized tests more extensively than did teachers with less training. A rather disturbing finding by these researchers was that the less well-trained teachers perceived themselves to be more knowledgeable about interpreting the results of tests than did the better trained teachers. In contrast, Green and Stager (1986-87) reported that the extent of teachers’ training in testing did not influence the frequency of their use of teacher-made tests, but they did find that the better (as compared to the less well-trained) teachers used somewhat more appropriate teacher-devised testing practices, such as the use of item analysis and test specification table procedures.

Not only classroom teachers but all educators tend to have had little or no training in educational measurement. Apparently, educators typically avoid measurement training when not required in their training program (Coffman, 1983; Schafer & Lissitz, 1987; Stiggins & Bridgeford, 1985). It has been suggested that educators may avoid measurement training because the training being provided is not designed to meet practical classroom demands (Airasian & Madaus, 1983; Stiggins & Bridgeford, 1985). In support of this explanation, Gullickson (1986a) identified major discrepancies between college measurement course topics and practicing teachers’ perceptions of what testing topics and skills are needed to successfully function in the classroom. He reported that classroom teachers place a heavy reliance on informal observations and direct pupil communications in making instructional decisions and perceive little need for statistical procedures. In contrast, Gullickson noted that preservice measurement instruction tends to focus upon paper-and-pencil measurement assessments and statistical analyses of data.

The findings from several other studies also suggest discrepancies between K-12 classroom teachers’ testing practices and their measurement training. Gullickson and Ellwein (1985) and Marso and Pigge (1988a) found that few practicing teachers use statistical analysis procedures in interpreting pupil test performance. Also, Kellaghan, Madaus, and Airasian (1982) reported that measurement training has resulted in little real impact upon teachers’ testing practices, and concluded that it is unlikely to do so until this training focuses on the actual demands of pupil assessment in classrooms. Finally, Gullickson and Hopkins (1987) reported evidence that many preservice measure-
ment professors themselves have limited measurement training and/or experience in the use of tests in K-12 classroom settings.

In addition to the major concerns about teachers having little or no preservice teacher training in testing and whether such training is appropriate, several researchers have reported that in-service teacher training in testing is almost nonexistent (Dorr-Bremme, 1983; Gullickson, 1984), and Marso and Pigge (1988a) found that neither teachers' ratings of their own testing proficiencies nor the quality of their teacher-made tests improved with the teachers' increased years of teaching experience. Further, what little in-service training teachers receive in testing and evaluation is commonly perceived by teachers as not being helpful. For example, Marso and Pigge (1987b) found that of all school experience factors assessed, first-year teachers were most dissatisfied with their in-service training. Furthermore, Stiggins (1988) has reported that teachers will seek in-service training designed to improve their tests and testing practices, but they will avoid in-service measurement training if it is perceived to be like that provided in preservice training.

In conclusion and as summarized in Table 3, it is apparent that K-12 classroom teachers are perceived by the educational and measurement communities to have limited testing knowledge and skills; that neither measurement consultative expertise nor in-service training in testing is generally available to teachers in their schools; that even basic testing support services, such as typing and duplication assistance, are not commonly available to teachers in a large number of schools; that a large portion of classroom teachers have had little or no formal preservice or in-service measurement training; and that much of the pupil assessment training available to teachers and teacher candidates is perceived by practicing teachers to be inappropriate for their classroom instruction settings.

Teachers' Testing Beliefs, Practices, and Attitudes

As noted previously, much of what we know about teachers' tests and testing practices has been obtained through studies using teacher self-report data gathering procedures. Few observational studies of teachers' testing practices or studies involving the direct analyses of teacher-constructed tests have been conducted. Consequently, we know little about what may be the true nature of classroom teachers' testing practices and the actual quality of their self-constructed tests (Stiggins, Conklin, & Bridgeford, 1986).
Table 3. Extent and Nature of the School and Measurement Communities’ Support for Teacher Testing as Suggested by Attitudes or Beliefs about Teachers and Teacher-Devised Testing, Extent of Teachers’ Training for Testing, and Availability of Resources for Classroom Testing

1. Just since 1990 have standards for classroom teachers’ testing competence been available; whereas standards for standardized testing have existed since the middle of the century.

2. The educational and measurement communities generally believe that teacher-constructed tests have a greater impact upon instruction and pupil learning in classrooms than do other types of tests.

3. The educational community and the measurement community perceive teachers, as well as many others in education, to have limited and inadequate classroom testing knowledge and skills.

4. Limited, if any testing expertise is available in most school buildings to assist and support teachers’ testing related responsibilities. Most educational training programs undergraduate and graduate, for K-12 administrative and teaching positions, with the exception of preparatory programs for guidance counselors, do not require training in testing and measurement.

5. Most educational administrators have little or no training in measurement and place limited emphasis on the management of testing and testing programs in the schools.

6. Building principals tend to believe that it is unnecessary to consult with testing specialists regarding testing and test development even in the development of district-wide tests.

7. Many K-12 classroom teachers have little or no formal training in tests and measurements. There are more teacher preparation institutions requiring no formal measurement training or just requiring training as part of another course than institutions requiring a complete course in tests and measurement for their teacher candidates.

8. Principals and teacher supervisors neither value nor encourage teacher use of technical testing skills such as use of item analysis, test specification tables, or test score statistical analysis procedures; teachers themselves do not deem these skills to be essential to the success of their pupil testing efforts.

9. As many as 20% of the standardized testing directors for school districts have no more training in formal tests and measurements than what is commonly expected of a classroom teacher.

10. Even basic support of teachers’ testing responsibilities such as typing and duplication services are not consistently available in approximately 50% of the schools. Computerized support services such as scoring, item analysis, etc. are available in just approximately 25% of the schools.

(continued...)
Table 3. (continued)

11. Teachers report that inservice training related to classroom testing and measurement is rarely if ever available in their schools. Limited evidence suggests that neither teachers' perceptions of their testing proficiencies nor the quality of their self-constructed tests improves with years of teaching experience.

12. Classroom teachers and a number of researchers have concluded that teacher preservice training in tests and measurements is not designed to meet the felt needs of K-12 classroom teachers. This may be part of the explanation for why preservice and inservice teachers, and other educators as well, generally do not participate in training in testing unless it is required of them.

13. School principals and teacher supervisors rate beginning teachers' proficiencies in tests and measurements somewhat lower than they rate beginning teachers' proficiencies in subject content or classroom management related skills.

14. The general educator community appears to convey the attitude that testing and measurement is a necessary but unpleasant process that does not deserve considerable attention or support.

15. Many college professors who instruct teacher candidates in educational measurement have limited formal training in measurement and/or limited experiences in the construction and use of tests and related measurement techniques in K-12 classrooms.

16. The measurement and education communities have conducted considerably less research on classroom teacher-devised testing as compared to the amount of research of standardized testing and of many other aspects of classroom instruction.

17. Limited research suggests that the availability of adequate school support and resources for testing positively influences teachers' attitude toward testing.

18. Neither school administrators nor teachers appear to perceive standardized testing in the schools to be primarily for their benefit (e.g., for administrative or instructional purposes).

19. Research evidence suggests that more teacher training in testing and evaluation result in more positive teacher attitude toward tests, more frequent use of classroom tests, more extensive use of standardized test scores, and somewhat more appropriate testing practices being used such as the use of item analysis and test specification table procedures.

Teachers' Classroom Testing Practices

It has been estimated that a typical pupil will take between 400 and 1,000 teacher-made tests before graduating from high school (Mehrens & Lehmann, 1987); that from 5% to 15% of a typical classroom day is devoted to some type of pupil assessment (Crooks, 1988; Haertel, 1986); and that teachers expend from 11% to 20% of a
typical work day on some aspect of pupil assessment, such as grading pupil work or preparing, administering, and scoring tests (Newman & Stallings, 1982; Stiggins, 1988). For example, in one study, teachers reported constructing an average of 54.6 formal paper-and-pencil tests in a typical school year (Marso & Pigge, 1988a) as part of their many and diverse pupil assessment activities.

Teachers rely primarily on their self-constructed tests, but many teachers frequently use publisher-constructed tests (textbook or workbook) tests as well in assessing their pupils. In one national sample of teachers, 95% reported using self-constructed tests and 77% reported using publisher-constructed tests (Dorr-Bremme, 1983). But regardless of the source of the test, teachers and pupils spend considerable classroom time and effort in testing activities (Fleming & Chambers, 1983).

Teachers' testing practices have been found to vary somewhat by grade level of instruction and by subject area content being assessed. At the upper grade levels, teachers rely more on teacher-constructed than publisher-constructed tests, express more concerns about the quality of pupil assessments, and use somewhat more test quality control procedures such as item analysis and checks on reliability than do teachers in the lower grades (Marso & Pigge, 1988a; Stiggins & Bridgeford, 1985). Primary grade teachers place more focus on pupil work samples than on testing; lower elementary grade teachers more frequently use worksheets and tests provided in publisher textbooks and workbooks than do other teachers; and upper grade and high school teachers predominantly use formal self-constructed tests in their assessment of pupils (Herman & Dorr-Bremme, 1982; Marso & Pigge, 1988a; Salmon-Cox, 1981).

Essay questions are very seldom used by classroom teachers at any grade level. Although infrequently used, essay questions are more frequently found in English, history, and social studies tests than in other subject area tests; and they are used more frequently in the upper grades than in the lower grades. Math and science teachers test their pupils more frequently than other subject area teachers, and they rely more heavily upon paper-and-pencil tests. Teachers in writing and speech classes are more likely to use direct observations and informal judgments than other teachers in assessing the progress of their pupils (Marso & Pigge, 1988a; Stiggins & Bridgeford, 1985).

Teachers in the upper grades tend to assign letter grades or marks based primarily on pupil test performance and daily work. In contrast, teachers in grades K-4 rely more on daily work and observations than on tests in assigning grades. Nevertheless, teacher-made
tests are considered to be at least one primary source of information about pupils for most teachers when assigning marks (Marso, 1986; Shulman, 1980).

Teachers rely more heavily on self-constructed tests than other types of tests in their instructional practices, and they typically report constructing from 50% to 75% of the test questions used on their tests. Teachers also use a variety of test items, with an average of 2.6 question types found on a typical teacher-devised test (Dorr-Bremme, 1983; Marso & Pigge, 1988a; Yeh, 1981).

Teachers most frequently use a combination of completion or short-response type questions in constructing their teacher-made tests, followed by the use of matching, multiple-choice, true-false, and essay type questions. When teachers are asked to rate the usefulness, adaptability, and fairness to pupils of the various question types, the question types are ranked in the following order: matching, completion, short-response, multiple-choice, true-false, and essay. Although essay tests are very infrequently used and perceived as not being very useful by most teachers, teachers believe that pupils study more for them than for objective tests, and that essay tests are more likely to measure higher cognitive levels than objective tests (Coffman, 1971; Marso, 1985).

Nearly all classroom teachers report that they provide pupils with feedback about their test performance following the administration of a classroom test, and typically they report spending about one half of a class period for that purpose. Teachers also report that pupils usually are very attentive and motivated during these test feedback sessions (Haertel, 1986). Once teachers construct test questions, they tend to reuse them without analysis and revision and, as noted previously, teachers report that they seldom use statistical procedures following the administration of a teacher-made test (Gullickson & Ellwein, 1985; Marso & Pigge, 1988c).

There are very few empirical studies revealing specifically how teachers use tests in their classroom instruction (Kuhs et al., 1985). Linn (1983), however, has described the linkage between classroom tests and instruction as consisting of these four basic features: the match between test items and the instructional objectives, test provision of feedback for pupil performance and teacher instruction, the “flag” role of tests in pointing out key content to be studied, and the use of tests to assist in assigning pupil letter grades.

A number of survey investigations of teachers’ testing practices have been conducted in the past decade. Generally, teachers report a heavy reliance on teacher-made tests in their day-to-day instruction;
in contrast, they report little reliance on standardized tests for making instructional decisions. Salmon-Cox (1981), after interviewing a sample of elementary teachers, reported that teachers made only minor use of the results from standardized tests in their classroom instruction, and Borg, Worthen, and Valcarce (1986) reported unfavorable and indifferent classroom teacher attitudes toward the use of standardized tests but a highly positive attitude toward the use of teacher-made tests. Stiggins and Bridgeford (1985) reported that classroom teachers use their self-constructed tests for pupil diagnosis, grouping, grading, evaluation, and reporting pupil progress in their classrooms. These latter researchers also reported that teachers place more reliance on teacher-made tests than on publisher-constructed tests (tests from workbooks, etc.), structured performance assessments, or spontaneous observations of pupils in making instructional decisions.

Dorr-Bremme (1983), following a survey of a national sample of school districts, revealed that the classroom assessments teachers rely on most heavily are characterized by immediate accessibility of scores, by an integration with teaching activities, and by a close tie between test questions and content taught. On each of these criteria, standardized tests are at a disadvantage, compared to teacher-made tests. At all grade levels and for all criteria assessed, teachers in a study reported by Hall, Carroll, and Comer (1988) attributed more value to teacher-prepared tests in making instructional decisions than standardized tests and as opposed to either district or state pupil minimum competency tests.

A persistent criticism of teachers is that they tend to overemphasize test scores (in particular standardized test scores) relative to other available information about pupils. Hall, Carroll, and Comer (1988) found, however, that classroom teachers consistently favored the results of their self-constructed tests over the results of standardized or state competency tests in making decisions. Further, they noted that teachers made decisions with a reasonable regard for the complex data requirements of classroom settings. Similarly, Lazar-Morrison, Polin, Moy, and Burry (1980) concluded that teachers place greater confidence in the results of their own judgments of pupil performance than in any formal tests. Furthermore, Stiggins and Bridgeford (1985) reported that teachers rely on a number of sources of information in making decisions about pupils and that teachers’ relative reliance on sources of pupil information is in the following order: teacher-made tests, standardized tests, structured performance assessments, and spontaneous observations.
Other research related to this concern about teachers' overreliance on test scores in making decisions about pupils also provides little support for this criticism of classroom teachers. Dorr-Bremme (1983) concluded that teachers bring several types of assessments to their decisions about pupils, and that they rely more on personal experiences and observations than upon test scores. Similarly, Salmon-Cox (1981) reported that high school teachers made very little use of standardized test scores in evaluating pupils; Shavelson, Cadwell, and Izu (1977) found that teachers gave due consideration to the reliability of data in making decisions about pupils; and Kellaghan, Madaus, and Airasian (1982) found that teachers can accurately predict pupil test performance and only use students' standardized test scores to corroborate their own judgments.

More specifically, the findings of the research related to teachers' use of test scores suggests that classroom teachers use scores to raise but not to lower their expectations of pupils. When teachers note a discrepancy between their perceptions of a pupil's ability and test scores, teachers ignore test scores when the scores suggest that less might be expected of a pupil, and teachers raise their expectations of a pupil when test scores suggest that more might be expected of a pupil (Airasian, Kellaghan, Madaus, & Pedulla, 1977).

Teachers' Attitudes and Beliefs About Testing

Although there is some inconsistency in the research findings about teachers' perceptions of their own testing ability, teachers typically rate the effectiveness of their training in testing somewhat below the training they received in other professional areas (Gullickson, 1984; Marso & Pigge, 1987a), rate their testing proficiencies somewhat lower than their proficiencies in other professional knowledge or skill areas (Marso & Pigge, 1987a), and express concern about their testing skills and believe that they could benefit from practical training in tests and measurements skills (Crooks, 1988; Haertel, 1986). Relatedly, first-year teachers rank the extent of their concerns about pupil evaluation and assessment above all other professional concerns except for their concerns about classroom management, pupil motivation, and coping with individual differences among pupils (Veenman, 1984).

Teachers commonly do not feel confident about their ability to write good test questions (Carter, 1984; Gullickson, 1985; Stiggins & Bridgeford, 1985) and are uncertain about how to improve their tests (Carter, 1984). Teachers report that they believe many of their questions and concerns about testing could be alleviated through
training (Carter, 1986). Conversely, several researchers have reported that teachers express confidence in their tests as well as in their overall testing knowledge and do not want more training in testing (Green & Stager, 1986-87).

This apparent conflict in findings, which suggests that teachers seemingly both desire but do not want more training in testing, may have been explained at least in part by Stiggins (1988). He noted that teachers do often express confidence in their overall tests and in their general testing knowledge. Conversely, he stated that teachers are uncertain about technical aspects of testing and that teachers do want practical help in improving their tests and their testing practices. What teachers do not want, he concluded, is more of the theoretical-impractical training typically associated with tests and measurement courses and workshops.

Two studies of teachers’ attitudes toward educational testing appear to be representative of teacher perceptions of tests and testing. Green and Stager (1986-87) surveyed 555 classroom teachers and reported that younger teachers are more skeptical of testing than older teachers; that upper grade teachers are more positive toward testing than lower grade teachers, who place more emphasis on classroom observations and informal pupil assessments; that teachers are positive toward teacher-made tests but tend to be negative or indifferent about standardized tests; that most teachers express interest in upgrading their testing skills; and that reported use of contemporary measurement practices (e.g., use of test specification tables and item analysis, etc.) was found to be somewhat related to more frequent pupil testing practices but not to attitude toward testing.

In a second study of teachers’ attitudes and beliefs about tests, Gullickson (1984) reported that teachers felt teacher-constructed tests result in increased pupil effort, influence pupil self-concept, create desirable competition among students, improve interaction among pupils, improve the classroom learning environment, better focus teaching, provide a good learning experience for pupils, motivate pupil study, and accurately reveal pupil progress. Further, Gullickson found that teachers believe frequent brief tests are more desirable than infrequent lengthy tests, school administrators encourage frequent testing of pupils, pupils prefer frequent tests, pupils try hard on tests, tests are an important instructional tool, tests need to be tied closely to instruction, tests help evaluate instruction, essay tests better assess pupil progress than objective items and measure at higher cognitive levels, tests should not be the sole determinant of grades, and tests are necessary to help justify grades to parents.
It may be that pupils reflect the attitudes of their teachers about tests, for students also feel that tests help them learn, and they too favor frequent testing. Pupils also report that teacher-made tests must be taken more seriously and are more difficult than standardized tests (Kulik & Kulik, 1988), and, like many teachers, some pupils feel that standardized tests are a waste of time (Stetz & Beck, 1981).

In summation, this review of teachers' testing practices, beliefs, and attitudes also revealed, as did the reviews presented previously, suggestions about teachers' testing knowledge and skills. Teachers expend considerable effort and time in fulfilling testing responsibilities in their classrooms; teachers schedule tests frequently followed by class discussions of pupil performance; teachers have concerns about, but also positive feelings about, the role of testing and pupil evaluation in the instructional process; and teachers have confidence in their classroom tests and their overall testing ability but recognize that they would benefit from practical training in testing. A summary of teachers' testing practices, beliefs, and attitudes is presented in Table 4.

Table 4. Teachers’ Testing Beliefs, Practices, and Attitudes

1. Teachers select and use assessment procedures that best fit their day to day instructional needs.
2. Teacher-made tests are perceived to better meet classroom instructional needs than do either standardized tests or state and school district pupil minimum competency tests.
3. Teachers believe that in order for test results to be of use to them tests must fit their instructional needs, must be of practical value, and must be immediately available.
4. Teachers believe that teacher-devised testing facilitates the classroom learning and teaching process.
5. Teachers believe, and feel that school administrators and pupils also believe, that teacher-made tests should be scheduled on a relatively frequent basis to promote pupil learning.
6. Teachers believe that teacher-made test assessments should closely mirror instruction provided.
7. Teachers believe that self-constructed assessments as compared to other assessments such as workbook and textbook tests generally better meet the instructional needs of their class.
8. Teachers believe that teacher-made tests generally have a positive impact upon pupils and their study-learning efforts.
9. Teachers believe that teacher-designed testing and the discussion of test results following the testing sessions are productive uses of classroom time.
10. Teachers believe that course content and pupil grade variations require somewhat different assessment devices and practices.
11. Teachers believe that test results should be supplemented with other sources of data such as observations and daily work when assigning grades and making decisions about pupils.
12. Teachers believe that daily experiences and teacher judgment are more reliable sources of data for making classroom and pupil related decisions than are isolated test scores.

(continued.....)
Table 4. (continued)

13. Teachers believe that where student learning is displayed in overt behaviors less reliance should be made of paper and pencil type tests.

14. Teachers believe that preservice training in tests and measurement provides them with adequate background concepts and principles but insufficiently prepares teachers for the successful integration of pupil assessments within the classroom instructional process.

15. Teachers believe that teacher-made tests can be relied on more than standardized tests and district or state competency tests for making decisions about individual pupils.

16. Teachers believe that teacher-made tests are useful in diagnosing pupils’ progress, making pupil grouping decisions, assigning pupil grades, and reporting the progress of pupils.

17. Teachers believe that essay tests as compared to objective tests are impractical and disliked by pupils but result in greater study efforts and usually measure at higher cognitive levels.

18. Teachers believe that they are less proficient in testing skills when compared to their proficiencies in other professional skill areas.

19. Teachers believe that testing, evaluation, and grading activities are among their more demanding and less pleasant classroom responsibilities.

20. Teachers believe that teacher-made test results aid teachers in justifying grades to pupils and parents.

21. Teachers believe that matching, short-response, completion, and multiple-choice questions are the more useable, efficient, and useful types of questions in contrast to the essay or true-false question types.

22. Teachers believe that testing and related assessment procedures, to be consistently used and useful in classrooms, must be efficient in time and energy demands of teachers and supportive of on-going classroom instructional activities.

23. Teachers believe that tests need to be administered fairly and efficiently and that testing periods should be monitored by teachers to prevent pupil cheating.

24. Teachers believe that test results can be interpreted and conveyed to pupils adequately without use of statistical analyses.

25. Teachers believe that a variety of question types should be used in classroom tests in order to be fair to pupils and to better complement various instructional objectives.

26. Teachers believe that teacher-made tests should contain questions that demand higher-order pupil thinking skills.

27. Teachers believe that technical aspects of classroom testing such as use of test specification tables, item analysis procedures, test score statistical analyses, estimates of test reliability, and use of question writing guidelines are of limited practical value.

28. Teachers generally report that they have deficiencies in testing and measurement, feel that their self-constructed tests could be improved, and would like inservice training in tests and measurements if this training were oriented toward practical classroom needs, but they tend to be confident about their general testing abilities and knowledge.

29. Teachers expend considerable class and work time and professional effort in testing and assessment activities, typically schedule formal tests once every two weeks or more often in most courses, construct on an average 54 formal tests each year, and construct most of their own test questions.

30. Most teachers place considerable reliance on information about pupils gathered through informal observations, daily communications, and daily work; teachers in the lower grades tend to rely more on these sources of information than on formal tests while middle and upper grade teachers tend to rely more on formal tests than upon informally gathered information.

31. Teachers believe that test scores must be interpreted and used within the context of all other information available about a pupil.

32. Teachers commonly express concerns about their pupil testing and evaluation responsibilities as well as about their class management and pupil motivation concerns.
DIRECT ASSESSMENTS OF TEACHERS' TESTING SKILLS AND KNOWLEDGE

As has been previously noted, very little research has been done involving the direct assessment of teachers' testing knowledge (Newman & Stallings, 1982). In this section, brief descriptions are presented of the findings from the very limited number of studies designed to directly assess teachers' testing knowledge, to rate the testing related proficiencies of teachers, or to directly assess teachers' test construction skills through analyses of their self-constructed tests.

Assessments of Teachers’ Testing Knowledge

Among the earliest efforts to directly assess teachers' testing knowledge was the study reported by Mayo (1967). He conducted a large-scale national study sponsored by the National Council on Measurement in Education and funded by the U.S. Office of Education. Two forms of the Measurement Competency Test were administered to 2,877 graduating seniors in 86 teacher-preparation institutions.

From an analysis of the data collected, Mayo concluded that teacher training practices at that time had not developed sufficiently the levels of measurement competency of beginning teachers to assure their success in meeting testing and evaluation responsibilities demanded in classroom instruction. Mayo recommended that preservice teacher measurement courses be improved; that a measurement course be compulsory for all teacher candidates; and that measurement courses have a practical focus, in order to better reveal to preservice teachers their need for measurement competencies and to increase their commitment to attaining these competencies.

Mayo's testing of graduating college seniors (1967) and his survey of testing professionals (1964) continue to be major reference points in the investigation of teachers' testing knowledge and skills, and the content of preservice measurement courses still reflects those topics deemed appropriate for the preparation of teachers by the testing professionals participating in the survey study. Providing further evidence of Mayo's continuing influence upon the measurement field, Newman and Stallings (1982) conducted what might be considered a follow-up of Mayo's study of teachers' testing knowledge. A battery of instruments patterned after Mayo's instruments, analyses of the content of several measurement textbooks, and a measurement item bank collected by the National Council on Measurement in Education were used by Newman and Stallings to assess the testing knowledge
6. TEACHERS' TESTING PRACTICES

of teachers who were employed in three large southern school districts. A total of 294 K-12 in-service teachers, identified through random selection procedures, completed this battery of assessment instruments. Some of the findings from this study that relate to the purposes of this chapter follow (the percentages in parentheses are comparable figures from the Mayo study):

1. Approximately 44% of the teachers in the sample had completed more training in measurement than one course, 33% (35%) had completed just one measurement course, about 6% (34%) took their measurement training as part of another course, and 13% (30%) had no formal measurement training.

2. The average percentage of questions answered correctly on the understanding of testing principles was 53.7% with teachers performing higher on general measurement principles than on technical aspects of testing.

3. As also was noted by Mayo, little difference in performance was found between trained teachers, with an average 54.6% correct response to the questions, and untrained teachers, with an average 48.0% correct response.

4. The teachers in the sample reported making about one half of their own tests and spent about 10% of their work time in testing activities.

5. The teachers in the sample reported greater use of objective than essay questions, with most to least frequent use of question types as follows: completion, multiple-choice, matching, true-false, short answer, calculation, and essay.

6. It was concluded from the collected data that there had been little change in the unacceptable level of teachers' testing knowledge since Mayo's study in 1967. Like Mayo, these researchers questioned the effectiveness of preservice teacher training in educational measurement.

Related, but less broadly based, studies tend to confirm the findings from the studies by Mayo and by Newman and Stallings. Carter (1986) found that teachers were unaware of item-writing faults or clues on a set of multiple-choice test questions, even though their seventh grade pupils were sufficiently testwise to use the faults in answering the questions. Hills (1977) reported that only 25% of the teachers in Florida show adequate measurement preparation and that just 10% to 20% can correctly answer basic questions on educational measurement principles. Impara, Divine, Bruce, Liverman, and Gay (1990) found that classroom teachers had difficulty in answering
questions related to scores derived from state-mandated achievement tests. These researchers also reported that those teachers with formal measurement training scored somewhat higher than those teachers without formal measurement training (a mean difference of about 1 on a 17-item test) and that interpretive information designed to accompany the score reports increased teacher performance on the questions. Without the interpretive information, 39% of the teachers answered fewer than 70% of the measurement questions correctly, whereas 10% of the teachers answered fewer than 70% of the measurement questions correctly with the information present.

In other studies, Carter (1984) found that language arts teachers were unable to recognize the particular skill being measured by test questions, that teachers took more time and found it more difficult to construct test questions functioning at higher cognitive levels, and that these teachers felt insecure about their knowledge of question-writing principles and previously had spent little time editing and revising test questions. Finally, the results of surveys of teachers' testing knowledge led Takeuchi (1977) and Infantino (1976) to conclude that teachers in California and New York had rather superficial knowledge of tests and measurement.

In summation, the findings from these studies utilizing direct assessments of teachers' tests and measurement knowledge levels suggest that teachers are not very knowledgeable about tests and measurement, and that neither preservice nor in-service training appears to be rectifying the situation. Many practicing teachers report having received no formal measurement training during preservice training, many teachers report having received only a unit of measurement training as a part of another preservice course, and most teachers report having received no school-sponsored in-service training or assistance in the development and use of tests in instruction (Dorr-Bremme, 1983).

Ratings of Teachers' Testing Proficiencies

Even though survey assessments of teachers' interests and skills commonly are used to help school administrators plan in-service instruction for teachers, just one study was located that had the major focus on the perceptual ratings of teachers' testing skills. Many other studies, however, collected and reported limited perceptual ratings of teachers' testing skills as secondary findings. The findings from these latter studies already have been reported in previous sections of this chapter.
Marso and Pigge (1991, 1989a, 1989b, 1989c, 1988c, 1987a) conducted a multifaceted statewide assessment of teachers' testing needs and proficiencies; findings from the various components of this study have been reported to audiences at different times and are referred to in different sections of this chapter. In this study, teachers, principals, and supervisors rated classroom teachers' proficiencies in 26 testing skills. Approximately 320 classroom teachers with 1 to 10 years of classroom teaching experience were asked to rate their current testing skill proficiencies, whereas the group of approximately 580 school principals and teacher supervisors were asked to rate the testing skill proficiencies of their typical beginning classroom teachers. Additionally, recently developed teacher-constructed formal tests were collected from the teachers and were assessed for question types used, cognitive functioning levels, construction quality, etc.

The 26 teacher testing competencies rated in this study are presented in Table 5 along with means derived from ratings completed on a 5-point Likert scale, with 5 being the highest proficiency rating. The rating means for this set of testing competencies are rank ordered for teachers and for school administrators. The supervisors' and principals' ratings were combined, as they were found to be highly correlated with one another. The teacher ratings of their testing proficiencies were found not to vary when classified by various levels of the teachers' years of teaching experience.

As can be noted in Table 5, the teachers rated their current testing skills higher than the administrators rated the testing skills of their typical beginning teachers. Even though the focus of the ratings differed between the two groups, the mean ratings of testing proficiencies for the two groups are relatively highly correlated, as can be noted by the similar mean rank orders for the two sets of rating means.

Both teachers and administrators rated teachers' proficiencies in writing several types of test questions relatively low as compared to other proficiencies. However, the testing skills associated with pupil grading and test scoring, selecting good test questions, and appropriately handling the format of tests were rated relatively high by both groups. When these teachers' tests were examined, however, it was found that the question-type writing skills rated highest by the teachers and administrators were the question types that violated more question-writing guidelines, and the question-writing skills rated lowest by the teachers and administrators were found to violate fewer accepted question-writing guidelines. In other words, a moderately high negative correlation was found between observed test
Table 5. Means and Ranks for Teachers’ Ratings of Their Current Proficiencies and Administrators’ Ratings of Beginning Teachers’ Proficiencies in Testing

<table>
<thead>
<tr>
<th>Testing Competencies or Skills</th>
<th>Teachers (N=313)</th>
<th>Administrators (N=580)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean*  Rank</td>
<td>Mean*  Rank</td>
</tr>
<tr>
<td>1. Writing good multiple-choice questions</td>
<td>3.64  19</td>
<td>2.99  9.5</td>
</tr>
<tr>
<td>2. Writing good completion questions</td>
<td>3.72  14.5</td>
<td>3.06  7</td>
</tr>
<tr>
<td>3. Writing good matching questions</td>
<td>3.81  9</td>
<td>3.10  6</td>
</tr>
<tr>
<td>4. Writing good true-false questions</td>
<td>3.58  20</td>
<td>2.99  9.5</td>
</tr>
<tr>
<td>5. Writing good essay questions</td>
<td>3.37  22</td>
<td>2.74  22</td>
</tr>
<tr>
<td>6. Scoring essay questions</td>
<td>3.21  24</td>
<td>2.67  24</td>
</tr>
<tr>
<td>7. Identifying good and poor questions for future tests</td>
<td>3.79  10</td>
<td>2.83  19</td>
</tr>
<tr>
<td>8. Writing questions in harmony with school and class goals</td>
<td>3.78  11</td>
<td>2.79  20</td>
</tr>
<tr>
<td>9. Stating objectives sufficiently clear to suggest test items</td>
<td>3.69  16</td>
<td>2.87  16</td>
</tr>
<tr>
<td>10. Writing test questions that demand higher thinking processes</td>
<td>3.52  21</td>
<td>2.55  25</td>
</tr>
<tr>
<td>11. Constructing tests that represent true student progress</td>
<td>3.65  18</td>
<td>2.78  21</td>
</tr>
<tr>
<td>12. Use of less formal assessments: checklists, ratings, etc.</td>
<td>3.28  23</td>
<td>2.86  17.5</td>
</tr>
<tr>
<td>13. Use of observations (visual) to assess and guide learning</td>
<td>3.72  14.5</td>
<td>2.95  11.5</td>
</tr>
<tr>
<td>14. Use of sociometric, guess who, and related techniques</td>
<td>2.88  26</td>
<td>2.72  23</td>
</tr>
<tr>
<td>15. Selecting good test questions from teacher manuals</td>
<td>3.93  5</td>
<td>3.13  5</td>
</tr>
<tr>
<td>16. Setting up readable, scorable, and attractive tests</td>
<td>3.88  7.5</td>
<td>3.02  8</td>
</tr>
<tr>
<td>17. Making tests reflect what is covered in text and class</td>
<td>4.23  3</td>
<td>3.19  3</td>
</tr>
<tr>
<td>18. Calculation of means, standard deviations, reliability, etc.</td>
<td>3.02  25</td>
<td>2.42  26</td>
</tr>
<tr>
<td>19. Interpreting test scores and student progress</td>
<td>3.75  13</td>
<td>2.88  14.5</td>
</tr>
<tr>
<td>20. Identifying individual and class strengths and weaknesses</td>
<td>3.91  6</td>
<td>2.95  11.5</td>
</tr>
<tr>
<td>21. Determining what needs to be retaught after tests</td>
<td>3.88  7.5</td>
<td>2.88  14.5</td>
</tr>
<tr>
<td>22. Use of tests and grades to positively influence learning</td>
<td>3.68  17</td>
<td>2.86  17.5</td>
</tr>
<tr>
<td>23. Calculating end of term grades from term work</td>
<td>4.25  2</td>
<td>3.43  1</td>
</tr>
<tr>
<td>24. Grading tests, papers, projects, homework, etc.</td>
<td>4.32  1</td>
<td>3.42  2</td>
</tr>
<tr>
<td>25. Deciding importance of tests, papers, etc. in grading</td>
<td>4.04  4</td>
<td>3.16  4</td>
</tr>
<tr>
<td>26. Deriving information from tests to guide students</td>
<td>3.97  12</td>
<td>2.91  13</td>
</tr>
</tbody>
</table>

*Means were derived from a 5-point Likert scale where 5 = high.
question-writing proficiencies and the perceived ratings of these testing proficiencies by the teachers and the administrators (Marso & Pigge, 1989c).

The classroom teachers in this study also rated the effectiveness of their preservice teacher training in tests and measurement lower than the effectiveness of their total teacher training experience, lower than the training received in their other education courses, and lower than the training in their arts and science courses. Similarly, the administrators rated the testing and measurement proficiencies of their typical beginning teachers lower than they rated beginning teachers' knowledge of their subject areas, lower than beginning teachers' other professional education proficiencies (e.g., instructional planning, handling discipline, etc.), and lower than beginning teachers' overall proficiencies as educators.

Assessments of Teacher-Made Tests

Rather surprisingly, very few studies of teachers' testing knowledge and skills have been conducted wherein direct analyses of teacher-made test samples have served as the major data-gathering procedure. One such study was reported by Fleming and Chambers (1983). They analyzed 342 teacher-made tests encompassing 8,800 test questions constructed by teachers assigned to several grade levels and subject areas in the Cleveland Public Schools. These tests and test questions were analyzed relative to Bloom's six cognitive functioning levels, question type use, subject content, grade level, and adherence to common question and format construction guidelines. Some of the more salient findings from this study follow:

1. Short-answer (including fill-in-the-blank) questions were most frequently used, followed by matching, multiple-choice, true-false (seldom used), and essay questions. Essay items were found very infrequently on any of these teachers' tests (about 1% of all questions).
2. Almost 80% of the questions found on the tests measured at the knowledge level. Approximately 94% of the questions on the junior high tests and 69% of the questions on all other tests examined were judged to be functioning at the knowledge level. The higher level functioning items, however, rather than being spread equally throughout all the tests, were found primarily on the math tests. Few questions on any tests were judged to measure pupils' ability to make applications.
3. Fewer than two thirds of the tests contained directions for all question types.
4. Questions were grouped by question type on all tests, but questions often were not numbered consecutively and in some cases were not numbered at all.

5. Suggestive of inadequate support services, many of the tests were handwritten, were poorly reproduced, and had pages overcrowded with content. Combined, these factors were deemed to make many of the tests almost illegible.

6. Commonly identified question-writing guideline violations included one or two word stems and illogical options in multiple-choice questions, matching items requiring fill-in-the-blank responses, and ambiguous short-answer response questions.

7. Most of the tests were approximately one or two pages in length and comprised approximately 35 questions, with fewer questions present on the tests for the lower grades.

In a second broadly based study of a sample of teacher-made tests, Marso and Pigge (1988a) analyzed 6,504 test questions contained within 455 question exercises (a group of questions of similar type on a test) found on 175 formal teacher-made tests, constructed by classroom teachers with 1 to 10 years of teaching experience who had completed a preservice tests and measurement course. These questions and tests were assessed for cognitive functioning level using Bloom’s six categories, violations of common test format and test question-writing guidelines, question types and numbers of questions used, subject content measure, years of teachers’ teaching experience, and test grade level, and by type of school setting (urban, rural, and suburban). Some of the more salient findings from this study follow:

1. Question type use varied by grade level and subject area content. Essay questions were very infrequently (about 1% of all questions) used by all teachers and were least used by elementary-level teachers, who more frequently used completion and multiple-choice questions than did secondary teachers. Problem questions (calculation tasks) were the predominant question form used by math teachers; science teachers most commonly used multiple-choice, matching, and short-response questions; and English teachers most commonly used short-response and matching questions.

2. Very few differences were noted in test construction practices or test construction quality when the tests were examined in terms of years of teachers’ teaching experience and type of school setting.
3. Matching exercises were found to be the most error-prone question type. Many question construction and test format construction guidelines were violated on many of the tests or test exercises, as shown in Tables 6 and 7.

4. Teachers reported preparing an average of 54.6 formal teacher-made tests each year, approximately 70% of the teachers scheduled a test once every 2 weeks or more frequently in a typical class, and over 50% of the teachers reported writing three fourths or more of the questions used on their tests.

5. The most frequently used question type used in the tests varied somewhat, depending upon whether the criterion used was total number of questions or most frequently used question type exercise. The question types used from highest to lowest frequency were short-response, matching, true-false, multiple-choice, problems, completion, interpretive exercises, and essay, as shown in Table 6.

6. As a total group of questions on all tests, 72% were judged to be functioning at the knowledge cognitive level. When examined by subject areas, this figure becomes more disturbing, as a large majority of the questions functioning beyond the knowledge level were contained just in the math and science tests. In other subject areas, the majority of the tests consisted of 90% to 100% questions judged to be functioning at the knowledge level.

7. Most teachers used a variety of test questions on their tests, with an average of 2.6 question types per test.

In another study involving the direct analysis of secondary math and science teacher-constructed tests, Oescher and Kirby (1990) analyzed 34 tests containing over 1,400 test questions and gathered the responses of 35 teachers to a teacher testing practices questionnaire. These teachers reported that summative evaluation is the dominant purpose of classroom testing in actual practice; that they wrote over 65% of the questions used on their tests; that they were confident in their ability to construct good tests; that they used instructional objectives to write items; that they discussed pupils' test results in class following an exam; and that they did not consistently use tables of test specification or item analysis procedures, or complete basic statistical analyses of their test scores such as the calculation of test score means. The direct analyses of these teachers' tests revealed that:

1. Format was in error on 70% of the tests (e.g., inadequate margins, spacing, etc.).
Table 6. Frequency of the Use, Construction Violations, and Bloom's Cognitive Functioning Levels of Question Exercise Types Found on 175 Teacher-Made Tests

<table>
<thead>
<tr>
<th>Item Types*</th>
<th>Number Items Reviewed</th>
<th>% Total Items Reviewed</th>
<th>No. of Exercises</th>
<th>% Tests with This Type</th>
<th>Total Talled Violations**</th>
<th>Mean Violations Per Exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching</td>
<td>1261</td>
<td>19</td>
<td>78</td>
<td>45</td>
<td>496</td>
<td>6.4</td>
</tr>
<tr>
<td>Completion</td>
<td>549</td>
<td>8</td>
<td>48</td>
<td>27</td>
<td>106</td>
<td>2.2</td>
</tr>
<tr>
<td>Essay</td>
<td>64</td>
<td>1</td>
<td>22</td>
<td>13</td>
<td>34</td>
<td>1.5</td>
</tr>
<tr>
<td>True/False</td>
<td>935</td>
<td>14</td>
<td>69</td>
<td>39</td>
<td>71</td>
<td>1.0</td>
</tr>
<tr>
<td>Multiple-Choice</td>
<td>1317</td>
<td>20</td>
<td>65</td>
<td>37</td>
<td>53</td>
<td>.8</td>
</tr>
<tr>
<td>Short Response</td>
<td>1093</td>
<td>17</td>
<td>89</td>
<td>51</td>
<td>61</td>
<td>.7</td>
</tr>
<tr>
<td>Problems</td>
<td>896</td>
<td>14</td>
<td>54</td>
<td>31</td>
<td>26</td>
<td>.5</td>
</tr>
<tr>
<td>Interpretive</td>
<td>362</td>
<td>6</td>
<td>30</td>
<td>17</td>
<td>6</td>
<td>.2</td>
</tr>
<tr>
<td>Unclassified</td>
<td>52</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>.1</td>
</tr>
<tr>
<td>Totals</td>
<td>6529</td>
<td>99</td>
<td>455</td>
<td></td>
<td>853</td>
<td>1.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple-Choice</td>
<td>1317</td>
<td>1123</td>
<td>7</td>
<td>112</td>
<td>73</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Matching</td>
<td>1261</td>
<td>1159</td>
<td>102</td>
<td>25</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Short Response</td>
<td>1093</td>
<td>830</td>
<td>235</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True/False</td>
<td>935</td>
<td>751</td>
<td>175</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Problems</td>
<td>896</td>
<td>35</td>
<td>59</td>
<td>798</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Completion</td>
<td>549</td>
<td>540</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Interpretive</td>
<td>362</td>
<td>199</td>
<td>118</td>
<td>40</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Essay</td>
<td>64</td>
<td>30</td>
<td>22</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Unclassified</td>
<td>52</td>
<td>28</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Totals</td>
<td>6529</td>
<td>4695</td>
<td>750</td>
<td>984</td>
<td>91</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Percent Each Cognitive Level: 72% for Knowl., 11% for Compr., 15% for Applic., 1% for Analysis, .001% for Synthesis, .001% for Eval.

*Selected item type definitions: Essay requires responses of paragraph or greater length; problem requires numerical calculation responses; interpretive requires answers to two or more questions following data presented (e.g., chart, table, map, poem, etc.); completion requires one- or two-word responses; and short response requires a phrase, a listing, or no more than one or two sentence responses.

**Violations tallied just once per item type exercise regardless of the times present.
Table 7. Question Exercise and Test Format Construction Errors Found on 175 Teacher-Made Tests: Frequencies, Percent of Errors, and Percent of Exercises with Error

a. Matching Exercises

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columns not titled</td>
<td>71</td>
<td>14</td>
<td>91</td>
</tr>
<tr>
<td>Not use &quot;once, more than once, or not all&quot; in directions to prevent elimination</td>
<td>69</td>
<td>14</td>
<td>88</td>
</tr>
<tr>
<td>Response column not ordered</td>
<td>60</td>
<td>12</td>
<td>77</td>
</tr>
<tr>
<td>Directions do not specify basis for match</td>
<td>55</td>
<td>11</td>
<td>71</td>
</tr>
<tr>
<td>Answering procedure not specified</td>
<td>52</td>
<td>10</td>
<td>67</td>
</tr>
<tr>
<td>Elimination due to equal numbers</td>
<td>46</td>
<td>9</td>
<td>59</td>
</tr>
<tr>
<td>Column(s) exceed 10 items</td>
<td>39</td>
<td>8</td>
<td>50</td>
</tr>
<tr>
<td>Materials not homogeneous</td>
<td>38</td>
<td>8</td>
<td>49</td>
</tr>
<tr>
<td>Premise not to left side</td>
<td>37</td>
<td>7</td>
<td>47</td>
</tr>
<tr>
<td>Numbers not to left and letters to right</td>
<td>13</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Exercise not contained on single page</td>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Requires responses to be written out</td>
<td>6</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Insufficient information in premises</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>496</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. Multiple Choice Exercises

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternates not in column(s) or rows</td>
<td>21</td>
<td>40</td>
<td>32</td>
</tr>
<tr>
<td>Incomplete stems</td>
<td>12</td>
<td>23</td>
<td>13</td>
</tr>
<tr>
<td>Negative words not emphasized or avoided</td>
<td>9</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>&quot;All or none above&quot; not appropriately used</td>
<td>5</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Needless repetition in alternatives</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Presence of specific determiners in alternates</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Verbal associations between alternate and stem</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Alternates overlap</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Needless phrases used</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Grammatical clues</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Distractors implausible</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Length clues</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>a and c, but not b, etc. used</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>53</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

c. Essay Exercises

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response expectations unclear, not labeled, etc.</td>
<td>14</td>
<td>41</td>
<td>64</td>
</tr>
<tr>
<td>Scoring points not realistically limited</td>
<td>7</td>
<td>21</td>
<td>32</td>
</tr>
<tr>
<td>Optional questions provided</td>
<td>5</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>Restricted question not provided</td>
<td>3</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>Ambiguous words used</td>
<td>2</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Opinion or feelings requested</td>
<td>2</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Question limited to simple listing response</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued....)
Table 7. (continued)

d. Problem Exercises

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items not sampling understanding or concepts, only calculations</td>
<td>20</td>
<td>70</td>
<td>37</td>
</tr>
<tr>
<td>Not range of easy to difficult problems</td>
<td>3</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Degree of accuracy not requested</td>
<td>2</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Nonindependent items</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Use of objective items when calculation preferable</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

e. Completion Exercises

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not complete interrogative sentence</td>
<td>32</td>
<td>30</td>
<td>67</td>
</tr>
<tr>
<td>Blanks in statements, &quot;puzzle&quot;</td>
<td>31</td>
<td>29</td>
<td>65</td>
</tr>
<tr>
<td>Textbook statements with words left out</td>
<td>18</td>
<td>17</td>
<td>38</td>
</tr>
<tr>
<td>More than single idea or answer called for</td>
<td>12</td>
<td>11</td>
<td>25</td>
</tr>
<tr>
<td>Question allows more than single answer</td>
<td>6</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Blank number clue</td>
<td>4</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Blank length clue</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Requests trivia versus significant idea</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Unstated degree of precision</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Lengthy, unnecessary words or phrases</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>106</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

f. True-False Exercises

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required to write response, time waste</td>
<td>20</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>Statements contain more than single idea</td>
<td>16</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>Negative statements used</td>
<td>15</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>Presence of specific determiner</td>
<td>8</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Statement not question, give away items</td>
<td>6</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Needless phrases present, too lengthy</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Imprecise statement, not always true or false</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Presence of length clue</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Opinion not attributed to source</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

g. Interpretive Exercises

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective response form not used</td>
<td>6</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Can be answered without data presented</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Errors present in response items</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Data presented unclear</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

(continued....)
Table 7. (continued)

h. Short Response

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item requires only listing</td>
<td>51</td>
<td>84</td>
<td>57</td>
</tr>
<tr>
<td>Response expectations ambiguous, not specified</td>
<td>7</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Unrealistically high scoring values assigned</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>61</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

i. Test Format

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absence of directions</td>
<td>82</td>
<td>29</td>
<td>47</td>
</tr>
<tr>
<td>Answering procedure unclear</td>
<td>61</td>
<td>22</td>
<td>35</td>
</tr>
<tr>
<td>Items not consecutively numbered</td>
<td>47</td>
<td>17</td>
<td>27</td>
</tr>
<tr>
<td>Inadequate margins</td>
<td>22</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Answer space not provided</td>
<td>21</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>No space between items</td>
<td>12</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Nonindependent items</td>
<td>11</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Different weighting of objective items</td>
<td>8</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Least time demanding types not first</td>
<td>7</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Similar item types not together</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>281</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

(Mean = 281 + 175 = 1.6)

%* Percentage of this specific error to all errors for this group (f=frequency of occurrence)

p* Percentage of all exercises of this item type with this specific error present

p' Percentage of tests with this type of specific format error

2. Directions were not present on 26% of the tests.

3. Over 60% of the questions were short-response questions, with multiple-choice, matching, and true-false comprising 20, 15, and 5% of all questions, respectively.

4. Just four essay questions were present among the more than 1,400 questions.

5. The teachers overestimated the number of their test items functioning beyond the knowledge level (Green, Halpin, & Halpin [1990] and Carter [1984] also noted this type of overestimation by teacher test writers). The teachers felt that about 25% of their questions measured beyond the knowledge and comprehension level, but judges determined the tests to contain an average of just 8% of all questions measuring beyond the knowledge and comprehension levels. Even few of the math test questions were judged to require pupils to apply knowledge of procedures to new situations.
6. All question types present on the tests were judged to violate several basic item-writing guidelines (e.g., 17 of 18 multiple-choice exercises contained major flaws, whereas short-response and true-false exercises were judged to be better constructed but still 50% of these question exercises contained construction flaws).

In other studies of less comprehensive samples of teacher-made tests, Billeh (1974) analyzed 33 science tests to determine cognitive functioning levels and reported that of all questions reviewed, 72% functioned at the knowledge level, 21% functioned at the comprehension level, and 7% functioned at the application level. The more experienced teachers in Billeh’s sample used more knowledge-level items, but no differences in the cognitive functioning levels of the tests were found when classified by grade level or by extent of teacher training. Black (1980) reported an analysis of 48 secondary-level science tests and found that the cognitive functioning levels of the tests varied within the science subject areas. Biology tests contained 94%, chemistry 66%, and physics 56% knowledge-level questions.

Similarly, Stiggins, Griswold, and Wikelund (1989) conducted interviews, class observations, and direct analyses of teacher-constructed tests of 36 K-12 classroom teachers. These teachers had been participating in in-service teacher training focused on school district-endorsed efforts to teach with a focus on the development of their pupils’ thinking skills. They found that all of these teachers’ self-constructed tests were composed of questions functioning 100% at the knowledge level except for the math tests. These researchers commented that it was easier to train teachers to teach with a focus on their pupils’ higher thinking levels than it was to train teachers to design tests to measure pupil achievement at these higher levels.

In summation, the review of studies of the ratings of teachers’ testing proficiencies, of the direct assessments of teachers’ testing knowledge, and of direct analyses of teacher-constructed tests have provided further suggestions about teachers’ testing knowledge, practices, and skills. School administrators and teachers themselves perceive teachers’ proficiencies in testing skills to be somewhat below their other professional proficiencies. The direct testing of teacher candidates’ and teachers’ knowledge about testing indicates that neither preservice nor in-service training in testing results in teachers being knowledgeable about basic testing concepts and principles. Direct analyses of samples of teacher-made tests reveal frequent violations of the most commonly accepted question-writing and test format-writing guidelines. Furthermore, teachers’ self-constructed
tests appear not to improve with increasing years of teaching experience. A summary of the more specific suggestions about teachers’ testing knowledge, practices, and skills derived from this review of studies of teachers’ testing proficiencies, knowledge, and tests are presented in Table 8.

Table 8. Teachers’ Testing Knowledge and Skills as Suggested by Perceptual Ratings of Their Testing Proficiencies, Tests of Their Knowledge, and Direct Analyses of Their Tests

1. Teachers’ more recent performance on measures of knowledge of classroom testing concepts and principles appears to be in the 50 percent correct range as was found in Mayo’s classic study in 1967. Some researchers have estimated that no more than 25 percent of K-12 classroom teachers can correctly answer basic questions on classroom measurement concepts and principles.

2. Teachers’ with formal training in tests and measurement perform better on measures of testing knowledge, but their scores typically exceed the scores of untrained teachers by just six to 10 percent.

3. Teachers tend to frequently use short-answer, completion, and matching question types which commonly measure at the lower cognitive demand levels. Multiple-choice questions are also commonly used; true-false are used less often; and essay questions are used very infrequently.

4. Teacher-constructed tests measure predominantly at the knowledge cognitive functioning level (approximately 70 to 100 percent range) with more higher level functioning items typically found on math and science tests and with tests in social studies and other subject areas functioning almost exclusively at the knowledge level.

5. Teachers display less knowledge and proficiency in technical aspects of testing (e.g., use of test specification tables, item analysis and statistical analysis procedures, etc.) and appear relatively unable to identify common item writing faults in test questions.

6. Analyses of teachers’ tests reveal very frequent violations of common question and format construction guidelines with matching exercises being found to be particularly error prone.

7. Principals and supervisors perceive beginning teachers and experienced teachers perceive themselves to have lower proficiencies in conducting simple statistical analyses of test scores, use of less formal data gathering procedures, writing questions demanding higher thinking skills, and use of sociometric techniques than in other testing proficiencies.

8. Teachers’, principals’, and supervisors’ ratings of teachers’ proficiencies in writing various test question types are highly but negatively correlated with directly observed frequencies of construction errors found in teacher-made tests.

9. The types of test questions used by teachers vary somewhat by subject area, content being assessed, and grade level of instruction.

10. Teachers have difficulty in correctly answering questions related to appropriate interpretations of scores commonly used in conveying pupil performance on standardized and state competency tests.

11. Many teacher-constructed tests are almost illegible due to poor typing or poor handwriting, lack of concern about format, and/or poor duplication quality.

12. Teacher-constructed tests typically contain approximately 35 questions with an average of 2.6 different question types being used and with questions grouped by question type.

(continued......)
Table 8. (continued)

13. Principals and supervisors rate beginning teachers’ testing proficiencies lower than beginning teachers’ proficiencies in other professional areas; practicing teachers also rate their testing proficiencies lower than they rate their professional proficiencies in other skill areas.

14. Many teacher-made tests contain incomplete, inadequate, or no directions.

15. Neither inservice training, if provided, nor increased years of teaching experience appear to improve either classroom teachers’ testing knowledge or their test construction skills as revealed by knowledge tests and by direct analyses of construction faults found on their self-constructed tests.

16. Teachers appear to value higher cognitive functioning questions on teacher-made tests, but they infrequently use such questions, tend to over-estimate the number of higher order questions used on their tests, and have difficulty identifying and writing test questions that function beyond the knowledge level.

17. Teachers appear to be unable to identify common test question construction guideline faults or violations on their tests and report spending little time editing or revising test questions. Some indirect evidence suggests that school principals and supervisors also are unable to distinguish between poorly and well written test question exercises.

18. Teachers, principals, and supervisors rate teachers’ grading related skill proficiencies higher than they rate teachers’ proficiencies in many other testing related skill areas.

19. Teachers, principals, and supervisors appear to agree rather highly with one another about the relative level of teachers’ proficiencies in various testing skills; they also agree with one another that teachers’ preservice preparation in testing is less adequate than their level of preparation in other areas of professional training.

Chapter Highlights and Recommendations

The purpose of this section is to highlight the general findings from the review of the measurement literature focusing on teachers’ testing knowledge and skills. A brief highlight of the findings from the research of each topic presented in this chapter is provided, followed by recommendations to the measurement profession on how it might better address the problem of the typical classroom teacher’s insufficient level of knowledge and skills related to testing and measurement. For a more extensive listing of summarization statements pertaining to the findings for the main topics reviewed in this chapter, the reader is referred to Tables 1 through 5 and Table 8.

Summary Highlights

The research literature available on classroom testing procedures, although predominantly comprised of studies conducted in university classrooms and characterized to some extent by inconsistent findings, suggests several possible generalizations related to teacher-devised testing practices. First, effectively designed classroom tests that are somewhat frequently scheduled have a generally positive
impact upon classroom learning. Second, the positive impact of testing upon pupil learning can be enhanced by announcing tests in advance; accurately describing the question types to be used and the content to be examined by the tests; closely matching test questions with instructional objectives; performing conscientious test administration and pupil monitoring during testing periods; and promptly returning the scored tests, accompanied by discussions of pupil performance on the tests and by appropriate reteaching of misunderstood concepts identified from an analysis of pupil performance on individual test questions.

Finally, characteristics of teacher-constructed tests that enhance either testing efficiency or pupil achievement are the following: inclusion of a variety of question types, each of which closely reflects the content being examined; inclusion of questions that function at a variety of cognitive levels placed in random difficulty order within question type categories; and inclusion of a sufficient number of questions to make full use of the amount of class time available and of appropriate difficulty to assure desired test reliability, as well as to challenge and reward pupil study efforts.

The educational and measurement communities’ support of K-12 classroom teacher-devised testing appears to be limited, uncoordinated, and of dubious merit. One clearly positive contribution, however, has been made by the measurement community in conveying its expectations of classroom teachers’ testing knowledge and skills through its 1990 standards for teacher competence in the educational assessment of pupils. But on the less positive side, educational measurement expertise is generally not available to K-12 classroom teachers in their schools. Also, educators’ attitudes toward testing and testing specialists borders on the negative. Many college of education deans, state legislators, and other educational leaders perceive a need for classroom teachers to have a higher level of testing knowledge, but collectively these groups tend to lend little or no support for either increased preservice or in-service teacher training in measurement. Many teachers, and most educators in general, receive little or no formal preservice training in tests and measurements, and much of the training provided is perceived to be narrow in scope and poorly designed to meet the instructional demands of the K-12 classrooms. Training in testing is frequently presented by college professors who themselves have limited measurement training and/or K-12 classroom experience in the construction and use of tests. Many practicing teachers have reported that in-service teacher training in tests and measurement does not exist. In many cases no
one in the local schools feels responsible for teacher training related to testing and test use, and most educators in the K-12 schools, regardless of their role in education, are not sufficiently knowledgeable about tests and testing to provide such training. Finally, basic support for teacher testing, such as typing and duplication services, and most basic computerized testing support services apparently are not consistently available to one half or more of the classroom teachers in many school districts.

Teachers, principals, and supervisors agree with one another that K-12 classroom teachers have a high need for testing knowledge and skills that clearly relate to and support the instructional process, but they are dubious about teachers’ need for more technical testing skills, such as the use of test specification tables and statistical analyses of test scores. Testing and related activities, such as assigning letter grades or marks in classrooms, appear to be perceived as necessary but unpleasant tasks by many in the educational community and, at best, these activities are deemed to be worthy of no more than grudging support. Further, the deluge of mandated testing in the schools in recent years may have accentuated rather than alleviated the problem of lack of availability of testing expertise in schools, the insufficient level of basic testing support services and resources, and the indifferent (if not negative) perception toward testing held by many in the educational community.

Classroom teachers generally value and are aware of the instructional benefits of teacher-instigated pupil testing, but they are far less positive about the value of district- and state-mandated pupil competency testing, and remain largely indifferent to the value of school district-sponsored standardized testing. Teachers perceive benefits of standardized testing to accrue primarily to others rather than to teachers in their school districts. Increasing numbers of research studies indicate that teachers use teacher-made tests in instructionally supportive ways, and tend to avoid potential negative labeling effects in their use of either teacher-made or standardized test scores in making decisions about pupils.

Several testing practices reported by K-12 classroom teachers and analyses of their self-constructed tests, however, suggest specific limitations in teachers’ testing skills and practices that somewhat mitigate against their generally positive instructional use of teacher-devised tests. More specifically, analyses of teachers’ testing practices and their self-constructed tests suggest the following: Test quality is generally poor and does not improve with teachers’ teaching experience, perhaps as the result of little or no in-service training in testing
and/or failure to use test improvement techniques such as question files, item analysis procedures, etc. Many teacher-made tests function almost exclusively at the knowledge cognitive level, perhaps due in part to factors such as teachers’ inability to construct and/or to distinguish between questions that function at higher and lower cognitive levels, insufficient teacher work time, and/or poor question type selection. Furthermore, the analyses of teacher-constructed tests reveal the presence of many construction faults, perhaps because teachers are unable to identify and revise these faults due to insufficient training, and/or because test question and test format construction guidelines are not readily available to them when they prepare their tests.

Recommendations to the Measurement Profession

The following recommendations need to be read, understood, and judged within the context of the following assumptions and conditions, as well as within the context of the findings from this review of the measurement literature pertaining to the testing knowledge, skills, and practices of K-12 classroom teachers. It should be noted that this review of the professional literature has revealed several findings positive to the measurement profession. For example, in recent years more research of teacher-constructed tests and their uses in K-12 classrooms has been conducted and is now appearing in the literature. Also, many in the measurement community, such as Richard Stiggins at the Northwestern Regional Educational Laboratory, are reporting instances of and methods for successfully increasing classroom teachers’ testing knowledge and skills.

It occurs to these writers, however, that it has been approximately a quarter of a century since the completion of Mayo’s landmark study (1967) revealing the inadequacies of classroom teachers’ testing knowledge and training. Many of his recommendations and findings remain as accurate and timely today as they were 25 years ago, and several researchers have concluded from recent studies that the extent of classroom teachers’ testing knowledge has changed little since the Mayo study. In light of this apparent lack of progress in improving teachers’ testing knowledge, the measurement profession probably needs to consider somewhat broader recommendations for alleviating these deficiencies than those typically found in the measurement literature, if the profession sincerely aspires to do more than describe the nature and extent of classroom teachers’ limited knowledge and training in tests and measurement.

The recommendations that conclude this chapter are primarily based upon an analysis and synthesis of the findings from the preced-
ing research reviews and the writings of noted leaders in the field. The following assumptions and conditions are presented to provide additional focus and a frame of reference for these recommendations:

- Teachers' self-constructed tests and associated testing practices in K-12 school classrooms are closely integrated with instruction, and demand considerable time and effort of teachers and pupils. Therefore, the provision of an increased level of support for these activities is likely not only to enhance the quality of these practices but is also likely to have a significant positive impact upon classroom teaching and pupil learning.

- Only those testing practices that are perceived to be practical, useful, and time efficient (if not timesaving) by K-12 classroom teachers are likely to be accepted and to persist in the schools.

- Teachers' testing knowledge and skills are inadequate, have not improved over the past 2 decades, and are not likely to improve in the future unless the measurement profession accepts the challenge of providing leadership to conduct long-term, coordinated, and cooperative efforts to address this inadequacy.

- To date, the professional measurement community's response to the inadequacy of teachers' testing knowledge and skills has been largely limited to a relatively undirected encouragement of better training practices, of further research of the problem, and of communications describing the problem.

- Many measurement professors and measurement specialists in other positions in the educational community are searching for meaningful research, training, and development opportunities. Their efforts and enthusiasm could greatly contribute to a concerted effort to address the problem of the inadequacy of teachers' testing knowledge, if these professionals could be provided with appropriate encouragement and direction.

- The current practical curricular, financial, and political constraints in higher education make it most unlikely that preservice teacher training in tests and measurement will be expanded to any great extent in the near future. Improvement in teachers' measurement training at this time can be addressed most effectively through increased and improved inservice teacher training, and through an emphasis upon more efficient and better focused preservice training in those institutions of higher learning where such training already exists.
Many individuals, professional groups, agencies, and members of the business community are currently interested in making schools more effective. They will likely make financial and human resource commitments to worthwhile efforts such as enhancing classroom instruction and pupil learning through improved teachers' tests and testing practices, if provided with encouragement and specific guidance in how to do so.

The recommendations directed primarily at the measurement profession for the purpose of ameliorating the inadequacy of teachers' testing knowledge and skills are:

1. The measurement profession, under the leadership of the National Council on Measurement in Education and the Buros Institute of Mental Measurements (in conjunction with other appropriate organizations of teachers and teacher educators, curriculum specialists, district superintendents, teacher supervisors, and building principals) should establish a task force to develop a broadly cooperative plan to address the continuing problem of classroom teachers' inadequate level of testing knowledge and skills, and the concomitant problem of insufficient expertise and resources in schools for the appropriate support of testing in the K-12 classrooms.

2. Instructional strategies and models for delivering both preservice and in-service teacher training in testing should be developed and field tested in order for the measurement profession to address seriously the problem of teachers' inadequate testing knowledge and skills. The focus of these products should be centered on practical classroom uses of tests and the development of specific test-writing and question-writing skills, as well as on the understanding of basic measurement concepts and principles.

3. The measurement profession, in conjunction with other appropriate professional education organizations, should develop and field-test a workshop or series of workshops designed to develop pupil assessment skills. Appropriate printed and other support materials should be designed to assist preservice and in-service teacher trainers in developing tests and measurement knowledge and skills, focusing upon appropriate and practical instructional uses of teacher-constructed tests in K-12 classroom settings.

4. The measurement profession, in conjunction with other appropriate professional educational organizations, should de-
velop and field-test a concise but effective classroom test analysis, test improvement, and test development training program with related printed and other support materials for use by supervisors of K-12 classroom teachers. At least one "certified" educational leader should be available in every school district to better supervise and support improved teacher-devised testing in K-12 schools.

5. The measurement profession, in conjunction with other appropriate professional educational organizations, should develop and refine through field testing a set of concise and illustrative test question-writing and test format construction guidelines, which should be made readily available to classroom teachers, teacher supervisors, and building principals.

6. The measurement profession, in conjunction with other appropriate professional educational organizations, should develop (or adopt existing) and field-test adaptable and user-friendly microcomputer software designed to provide test scoring, item analysis, estimates of test reliability, and related computerized testing support services for teachers in every school building.

7. The measurement profession, in conjunction with other appropriate professional educational organizations, should instigate the development of a program designed to make available in each school building basic teacher testing responsibility support services, such as typing, duplication, computerized testing support service operations, etc. Human resources might be arranged through parent-teacher associations in conjunction with internship arrangements from high school business education or future teacher programs, etc.

8. The measurement profession, in conjunction with other appropriate professional educational organizations, should develop a mechanism—perhaps an agreement by all major textbook publishers—to add a small amount to the selling price of each textbook sold. This would generate financial support for creating test question-writing services to assure a substantial improvement in the number and quality of test questions made available in instructional manuals, workbooks, and chapter tests to accompany all major textbooks used at all educational levels.

9. More studies involving direct analysis of samples of teachers’ self-constructed tests should be conducted to determine more precisely the nature and quality of these measurement instru-
ments, and to provide further insight into how more effectively to assist teachers in improving their tests. For example, we need to know more about how to encourage teachers to construct structurally sound questions that function at higher cognitive levels, and we need to know more about the long-term impact that such questions have on pupil study, thinking, and learning.

10. More studies should be conducted to provide further insight into the nature and extent of instructional uses of teacher-designed tests. The existing research literature indicates that current tests and measurement training does not adequately address the practical, instructionally integrated uses made of tests by teachers in actual classrooms. We need to know more specifically what these practices are and how this knowledge can be translated into more appropriate preservice and inservice teacher training activities.

REFERENCES


6. TEACHERS’ TESTING PRACTICES


Newman, D. C., & Stallings, W. M. (1982). *Teacher competency in classroom testing, measurement preparation, and classroom testing*


One of the most common activities in which teachers engage is assessment of students. Stiggins and Conklin (1988) estimated that teachers spend as much as a third of their professional time in assessment-related activities. Although teachers need a variety of observational and problem-solving skills for effective classroom assessment (see chapter 2 by Richard Stiggins in this volume), a substantial portion of classroom assessment activities draws upon teachers’ skills in testing and measurement. If they do not have a firm understanding of basic principles of measurement, teachers are more likely to engage in unsatisfactory assessment practices. Hence, a necessary (though by no means sufficient) requirement for effective classroom assessment is that teachers be skilled in measurement.

By and large, the measurement demands being placed on the classroom teacher appear to be increasing, both in amount and sophistication. Curriculum-based assessment, which requires frequent testing of students, is being implemented in an increasing number of schools. Criterion-referenced (i.e., mastery) testing, for which proper use requires measurement knowledge and skills that are substantially different from those needed for norm-referenced
testing, is becoming more common. Moreover, recent technical advances in measurement, such as item response theory, are being implemented with increasing frequency in school-based testing programs.

The research base on teacher training in measurement has indicated cause for concern. In the most comprehensive study of this issue to date, Schafer and Lissitz (1987) surveyed the measurement training practices of the American Colleges of Teacher Education (AACTE) member institutions. They found that less than half of the teacher education programs required a formal course in testing and measurement for graduation. Moreover, this is not a newly identified problem. Noll (1955) reported that only 21% of a sample of teacher education programs required a course in measurement. He concluded that prospective teachers’ training in testing and measurement is “almost certainly inadequate to prepare them to function effectively in an area so essential to their success as teachers. The situation should be a real matter of concern to all engaged in the work of educating teachers” (p. 90). Apparently, the level of concern has not grown too greatly, given the findings of Schafer and Lissitz (1987) that most teacher education programs do not require a course in measurement.

Why has the measurement training of teachers remained underemphasized? To a large extent, the curricula in teacher education programs are determined by state requirements for certification. Little pressure has apparently been exerted on programs by state departments of education for more extensive measurement training. Wolmut (1988) found that only 20% of the states either require a measurement course or list specific measurement-related content requirements for the certification of teachers.

How do teachers feel about this discrepancy between their measurement training and the measurement demands of their jobs? The small amount of research in this area suggests that teachers feel that they have sufficient measurement skills. Gullickson (1984) surveyed 391 teachers regarding their measurement-related attitudes. He concluded that (a) teachers perceived their knowledge of testing and measurement as being adequate and (b) most teachers believe that they have learned about testing and measurement through their classroom experiences.

The purpose of the current investigation was twofold. First, the amount of formal measurement training provided by each of the teacher education programs in Nebraska was studied. Second, a sample of Nebraska school teachers was surveyed to identify relation-
ships between the amounts of formal measurement training of practicing teachers and (a) their beliefs about the adequacy of their training, (b) their perceived importance of measurement coursework, (c) factors influencing their measurement knowledge, and (d) their own perceived abilities in measurement.

Part of the motivation for conducting this study concerns a common attitude that often seems to be held regarding educational problems. That is, although teacher educators will acknowledge that there is a particular problem in education, they do not feel that the problem is prevalent in their state. Because the lack of measurement training in teachers has been known about for decades, and yet has led to little change in teacher education practice, we suspect that an “it really isn’t a problem here” attitude may have contributed to the small degree of change in teacher measurement training that has occurred since Noll’s (1955) study.

The state of Nebraska was chosen for this study primarily because of convenience, and also because it happened to be the setting for the Buros-Nebraska Symposium on Measurement and Testing at which these results were presented. Nebraska typically fares quite well in comparisons with other states in terms of student achievement. What we found in Nebraska, however, we feel is representative of most, if not all, other states in the U.S. Hence, readers of this chapter should keep in mind that the findings in Nebraska are likely to be indicative of their states.

MEASUREMENT TRAINING IN NEBRASKA

Inquiries were made to the 15 Nebraska universities and colleges that offer teacher preparation programs. Information about required coursework in measurement, including course names and numbers, credit hours, percent of instructional time, and topics covered was gathered via telephone conversations and through course catalogues. All 15 programs devoted some instructional time to measurement topics. The topics that were typically included in instruction were (a) statistics, (b) reliability, (c) validity, (d) test construction, including information about item types and item analysis, (e) uses of standardized tests, (f) interpretation of standardized test scores, (g) standardized test norms, and (h) use of standard scores on standardized tests. A brief summary describing course offerings at these institutions follows.

The University of Nebraska-Lincoln (UNL) has the largest teacher preparation program in the state, graduating approximately 450 stu-
dents per year. There is one required course that covers measurement topics, offered through the Educational Psychology department. Approximately one third of this one-semester, three-credit-hour course is devoted to measurement topics. The topics routinely covered are reliability, validity, test construction, and standardized tests.

The University of Nebraska at Omaha (UNO) has the second largest teacher preparation program in the state, graduating approximately 400 students per year. Students are required to take Human Growth and Learning, offered through the Education department, which covers standardized testing as well as numerous other topics. Students are also required to take a course in Instructional Systems, which is partially devoted to the coverage of measurement topics. Students receive instruction in objectives, teacher-made tests, grading, and alternative forms of assessment such as student products and checklists. In addition, measurement topics are covered in the methods courses.

University of Nebraska at Kearney (UNK) teacher preparation program graduates approximately 150 students per year. Students are required to take Learning and Evaluation, which is a one-semester, three-credit-hour course in the Education department with approximately one fourth of the instructional time devoted to measurement topics. These topics include (a) selecting and/or designing tests, (b) utilizing information from tests, and (c) using and interpreting standardized tests.

Concordia Teachers College graduates approximately 150 education students per year. All students, except Elementary Education majors, are required to take an Educational Measurements course in the Psychology department. This course is a one-semester, three-credit-hour course focusing exclusively on measurement topics. The goal of this class is to teach students to administer and interpret a variety of tests, including norm-referenced, criterion-referenced, informal, and functional.

The teacher preparation program at Wayne State College also graduates approximately 150 students per year. Tests and Measurement, in the Education department, is an optional course offered for three credit hours. The entire course is devoted to measurement topics including (a) historical background, (b) objectives, (c) test construction, (d) anecdotal records, (e) measurement of attitudes and social behavior, (f) statistics, (g) validity, (h) reliability, and (i) standardized tests.

Chadron State College graduates approximately 100 students per year from their teacher preparation program. The program at Chadron
7. MEASUREMENT TRAINING

State requires one Education department course on measurement theory, Elementary/Middle School Tests and Measurements. This one-semester, one-credit-hour course is devoted to measurement topics including reliability, validity, test construction, and standardized tests. There is a heavy emphasis on test construction in this course including a discussion of item types and statistical analysis of tests.

The approximately 100 graduates of the teacher preparation program at Peru State College have the option of including an Education department course, Tests and Measurements, as part of their educational program. This one-semester, two-credit-hour course is devoted entirely to measurement topics. These topics include (a) issues, (b) objectives, (c) teacher-made tests, (d) anecdotal records, (e) statistics, (f) reliability, (g) validity, and (h) standardized testing.

Creighton University’s teacher preparation program requires their approximately 80 graduates per year to take Educational Psychology, offered through the Education department. Approximately one third of this one-semester, three-credit-hour course is devoted to measurement topics. These topics include reliability, validity, teacher-made tests, norms, standard scores, and standardized tests. In addition, test construction is covered in the methods courses offered through this program.

Hastings College graduates approximately 45 students per year from their teacher preparation program. These students receive approximately 10 hours of instruction on measurement topics in methods courses and in their senior seminar. These topics include reliability, validity, test construction, and standardized tests.

The approximately 30 students who graduate yearly from Midland Lutheran College’s teacher preparation program receive approximately 8 hours of instruction on measurement topics. This instruction is offered as part of the curriculum and general methods courses. The topics that are covered include reliability, validity, and test construction.

Dana College graduates approximately 25 students per year from the teacher preparation program. These students are required to take Tests and Measurement, offered through the Education department. This course is a one-semester, three-credit-hour course devoted entirely to measurement topics, including (a) reliability, (b) validity, (c) standardized tests, (d) test construction, (e) evaluation instruments, (f) observations, (g) checklists, (h) student products, and (i) assessing learning styles.

Doane College’s teacher preparation program graduates approximately 20 students per year. These students are required to take an
Education department course, Measurement and Evaluation. This is an 8-week mini-course that meets for 3 hours per week. The topics covered in this course are (a) objectives, (b) reliability, (c) validity, (d) test construction, (e) grading, (f) evaluation of special needs, and (g) observational techniques. Standardized tests are discussed in other required courses.

Union College’s teacher preparation program also graduates approximately 20 students per year. These students are required to take Learning Theory and Measurement, which is offered through the Education department. This course includes approximately 9 hours of instruction on measurement topics. Students are taught how to interpret standardized test scores and construct classroom tests. They also learn about measuring individual differences, with a particular focus on intelligence.

Students graduating from the teacher preparation program at the College of Saint Mary are required to take Educational Psychology and Measurement, offered through the Education department. This is a one-semester, four-credit-hour course that is partially devoted to measurement topics. The curriculum includes a discussion of evaluative tools and standardized tests.

Nebraska Wesleyan University’s teacher preparation program offers several required Education department courses that focus on measurement topics. Educational Measurements (Secondary) is a one-semester, three-credit-hour course that covers teacher-made and standardized tests. Secondary - Educational Measurements: Directed Study and Special Education - Educational Measurements: Directed Study are both one-semester courses offered for 2 credit hours. These courses are tailored to fit with the programs of individual students.

Table 1 provides summary information concerning the above mentioned programs. This table shows that 73% of the teacher education programs in the state of Nebraska require their students to take less than one full course in measurement. This 73% includes two of the largest programs in the state, the University of Nebraska-Lincoln and the University of Nebraska at Omaha. These two programs graduate approximately 49% of the students enrolled in teacher preparation programs in the state.

These results are consistent with the results of Schafer and Lissitz (1987), who found that less than half of the teacher education programs required a formal course in testing and measurement for graduation.
### Table 1. Summary of Teacher Preparation Programs in Nebraska

<table>
<thead>
<tr>
<th>Institution</th>
<th>Approximate Percent of State Graduates</th>
<th>Measurement Training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Entire Course</td>
</tr>
<tr>
<td>UNL</td>
<td>26%</td>
<td>X</td>
</tr>
<tr>
<td>UNO</td>
<td>23%</td>
<td>X</td>
</tr>
<tr>
<td>UNK</td>
<td>9%</td>
<td>X</td>
</tr>
<tr>
<td>Concordia</td>
<td>9%</td>
<td>X</td>
</tr>
<tr>
<td>Wayne</td>
<td>9%</td>
<td>X</td>
</tr>
<tr>
<td>Chadron</td>
<td>5%</td>
<td>X</td>
</tr>
<tr>
<td>Peru</td>
<td>5%</td>
<td>X</td>
</tr>
<tr>
<td>Creighton</td>
<td>5%</td>
<td>X</td>
</tr>
<tr>
<td>Hastings</td>
<td>3%</td>
<td>X</td>
</tr>
<tr>
<td>Midland</td>
<td>2%</td>
<td>X</td>
</tr>
<tr>
<td>Dana</td>
<td>2%</td>
<td>X</td>
</tr>
<tr>
<td>Doane</td>
<td>1%</td>
<td>X</td>
</tr>
<tr>
<td>Union</td>
<td>1%</td>
<td>X</td>
</tr>
<tr>
<td>St. Mary</td>
<td>unknown</td>
<td>X</td>
</tr>
<tr>
<td>Wesleyan</td>
<td>unknown</td>
<td>X</td>
</tr>
</tbody>
</table>

**TEACHER BELIEFS AND ATTITUDES**

A 13-item survey was developed to gather information about teachers’ (a) demographic characteristics, (b) training in testing and measurement at the preservice, inservice, and graduate levels, (c) feelings about the adequacy of their undergraduate training in measurement and testing, (d) influences on their knowledge of measure-
ment, and (e) self-assessments regarding their abilities in various aspects of measurement. It was designed to be completed in a short period of time; pilot testing showed that most teachers needed less than 10 minutes to complete the two-page survey.

Sample and Procedure

There were 825 surveys sent to the teachers in two Nebraska school districts, one of which was predominantly rural and the other predominantly urban. Participation in the study was voluntary. The surveys were deposited in the teachers' mailboxes along with a cover letter providing a brief explanation of the nature of the study and encouraging teacher participation. At most of the schools, a drop-off box was placed in the main office for completed surveys. Several days prior to the deadline for returning the surveys, a brief memo was sent to the teachers reminding them of the upcoming deadline, if they chose to participate.

A total of 397 completed surveys were returned by teachers, which corresponded to a return rate of 48%. The breakdown of respondents, by level of school taught, was as follows: elementary school, 41%; junior high school, 34%; high school, 25%. These percentages were consistent with the distribution of teachers at each level in the two districts studied. The respondents reported an average of 15.35 years of teaching experience. Eighty percent of the respondents reported receiving their undergraduate training at one of the teacher education programs in the state of Nebraska.

Survey Results

The measurement training of the respondents was quite varied; 15% reported that they had received no coursework in measurement, 51% reported that part of one course was devoted to measurement, 25% reported taking one entire measurement course, and 9% reported taking two or more measurement courses. These results are consistent with those found in the nationwide survey of Schafer and Lissitz (1987). In Tables 2-4 below, it was useful to separate the sample of respondents into two subgroups: those with less than one course in measurement (66%), and those with one or more courses (34%).

One of the survey questions concerned respondents’ feelings about the measurement training that they had received as an undergraduate. Table 2 shows that, for the entire sample of respondents, almost half (47%) felt that their training was somewhat or very inadequate. Moreover, there was a clear discrepancy between the
feelings of the training subgroups. A clear minority (18%) of those with one or more courses reported that their measurement training was at least somewhat inadequate, whereas a clear majority (64%) of those with less than one course felt that their training was adequate.

Table 2 also contains information about whether or not respondents had received any additional formal measurement training, either in graduate courses or measurement-related inservice training. Only about a third of the respondents had taken a graduate course in measurement, and only about a fifth of the sample reported measurement-related inservice training. In terms of the training subgroups, however, additional training was markedly different. Teachers with one or more undergraduate measurement courses reported both substantially more graduate coursework and more inservice training than those teachers with less than one undergraduate course. Hence, even though teachers with less than one undergraduate course reported greater dissatisfaction with their undergraduate training, they were less likely to acquire formal measurement training after completion of their undergraduate studies.

Where, then, do teachers learn about testing and measurement? Table 3 displays the results of a survey question concerning the factors that had the greatest impact on the respondents' measurement knowledge. For the total group of respondents, a majority of the teachers cited trial and error learning in the classroom as having the greatest impact, with college/university coursework ranking a distant second and one’s own reading third. The rank orders of the three categories are the same for each of the training groups, but the training groups showed differences in the relative percentages choosing each category. Formal coursework had a much stronger relative effect on those respondents with at least one measurement course. For respondents with less than one course, 80% identified non-coursework factors as having the greatest influence on their measurement knowledge.

Respondent agreement with a statement regarding the importance of measurement skills to teachers being perceived as professionals is shown in Table 4. Approximately three-quarters of the respondents agreed or strongly agreed with the statement. There did not appear to be a substantial difference between the training subgroups in terms of their ratings of the statement.

Respondents were also asked to rate their own abilities in a variety of areas of measurement. Table 5 shows the ratings of the total sample for each area. For each area, a “Not Applicable” choice was provided for those respondents who felt that the measurement area
<table>
<thead>
<tr>
<th>Group</th>
<th>Evaluation of Undergraduate Measurement Training</th>
<th>Graduate Courses in Measurement?</th>
<th>Inservice Courses In Measurement?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Very Inadequate</td>
<td>Somewhat Inadequate</td>
<td>Somewhat Adequate</td>
</tr>
<tr>
<td>All Respondents</td>
<td>24%</td>
<td>23%</td>
<td>35%</td>
</tr>
<tr>
<td>Those with One or More Courses</td>
<td>12%</td>
<td>6%</td>
<td>46%</td>
</tr>
<tr>
<td>Those with Less Than One Course</td>
<td>31%</td>
<td>33%</td>
<td>29%</td>
</tr>
</tbody>
</table>

Table 2. Evaluation of Undergraduate Measurement Training and Amount of Post-Graduate Training Attained
Table 3. Factors Affecting Respondents' Knowledge of Testing and Measurement

<table>
<thead>
<tr>
<th>Group</th>
<th>Greatest Effect on Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>College/University Coursework</td>
</tr>
<tr>
<td>All Respondents</td>
<td>28%</td>
</tr>
<tr>
<td>Those with One or More Courses</td>
<td>42%</td>
</tr>
<tr>
<td>Those with Less Than One Course</td>
<td>20%</td>
</tr>
</tbody>
</table>

Table 4. Respondent Agreement with the Statement, “In Order for Teachers to be Perceived as Professionals, it is Important That They Possess Strong Skills in Technical Areas Such as Testing and Measurement”

<table>
<thead>
<tr>
<th>Group</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Respondents</td>
<td>16%</td>
<td>57%</td>
<td>21%</td>
<td>5%</td>
</tr>
<tr>
<td>Those with One or More Courses</td>
<td>20%</td>
<td>51%</td>
<td>24%</td>
<td>5%</td>
</tr>
<tr>
<td>Those with Less Than One Course</td>
<td>14%</td>
<td>61%</td>
<td>18%</td>
<td>6%</td>
</tr>
</tbody>
</table>

did not apply to their jobs. The ratings were highly consistent across measurement areas; the majority of the respondents felt that their abilities were good or very good. Very few respondents rated their abilities as very poor. The respondents were next asked to rate the importance of the same measurement areas to their jobs. The ratings given by the total sample are displayed in Table 6. The respondents rated most of the areas as important or very important. The ratings of the two areas concerning the administration and interpretation of standardized tests, although still fairly high, were markedly lower
than those given to the other areas. The results reported in Tables 5 and 6 were highly similar across training groups; hence, they were not broken down by those groups.

Discussion and Conclusions

Taken together, the analysis of the Nebraska teacher education programs and the findings from the teacher survey describe a situation that should be of concern to teacher educators. Approximately 84% of the graduates of Nebraska teacher education programs currently receive less than one full course in measurement. In the teacher survey, two-thirds of the respondents reported that they received less than one undergraduate course in measurement. Moreover, teachers with less than one undergraduate course do not seem to feel that there is a deficit in their training and seek measurement instruction via graduate courses or inservice training. Most teachers rated measurement skills as an important component of professionalism in teaching, and they tended to rate their own measurement skills highly. The source of these skills was reported to be largely trial-and-error learning in the classroom.

Table 5. Respondents’ Ratings of Their Own Abilities in Various Measurement-Related Areas

<table>
<thead>
<tr>
<th>Area</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Very Poor</td>
</tr>
<tr>
<td>Constructing and improving classroom tests</td>
<td>0%</td>
</tr>
<tr>
<td>Administering standardized tests to students</td>
<td>0%</td>
</tr>
<tr>
<td>Interpreting scores from classroom tests</td>
<td>0%</td>
</tr>
<tr>
<td>Interpreting scores from standardized tests</td>
<td>0%</td>
</tr>
<tr>
<td>Understanding of test reliability and validity</td>
<td>2%</td>
</tr>
<tr>
<td>Explaining the meaning of test scores to others (e.g., parents)</td>
<td>2%</td>
</tr>
<tr>
<td>Scoring (grading) classroom tests</td>
<td>0%</td>
</tr>
</tbody>
</table>
Because teachers do not seem to feel that their measurement skills are inadequate, it is tempting to characterize any problems caused by limited undergraduate training as self-correcting. That is, through their experiences in the classroom, teachers eventually acquire measurement skills on their own. Unfortunately, the idea that the problem corrects itself is unsupported by empirical research, which has indicated that there are widespread deficits in the measurement skills evidenced by practicing teachers (Carter, 1984; Fleming & Chambers, 1983; Newman & Stallings, 1982; Stiggins and Bridgeford, 1985).

Why, then, do teachers rate their measurement skills so highly, in light of evidence to the contrary? An easy answer is that teachers do not want to admit to their deficiencies. We feel, however, that most teachers genuinely believe that their skills are adequate. The problem may instead lie in the culture of the schools. Aspiring and practicing teachers continually receive messages that measurement skills are not very important. This socialization begins in the teacher education programs, where required instruction in measurement is minimal. Moreover, college and university faculty, most of whom have no
measurement training, often provide poor models for how to measure student achievement. After graduation, a new teacher enters an environment in which the other teachers are generally poorly trained in measurement, as are the school administrators. Teachers are not held accountable for having reliable and valid measurements of their students. Many teachers associate measurement with standardized testing, which has elicited strongly negative attitudes from teachers, administrators, parents, and students. Hence, it is relatively easy to imagine that many teachers undervalue measurement skills. They have been trained and work in environments in which no one has ever explained how such skills could allow them to be more effective decision makers in the classroom and make better inferences about their students.

Teachers may believe that their measurement abilities are strong, and they receive little feedback to the contrary. If two teachers, one strong and one weak in measurement skills, each develop and administer a test to their students, each teacher will acquire a set of test scores that does not appear to differ from the other set. As long as each teacher believes that his or her test is reliable and valid, the two teachers may be equally comfortable with the resultant scores. There appears to be no mechanism in the schools to provide feedback to teachers on the quality of their measurements and assessments. In the absence of feedback, beliefs may play a major role.

Another potential explanation for the lack of measurement training is that teachers may find such training to be anxiety provoking. As it is typically taught, much of the content of a measurement course involves the understanding and proper use of formulas. Such course content can produce mathematics anxiety similar to that experienced by students in statistics courses. In addition, anxiety about measurement might be associated with negative testing experiences that teachers may have had when they were students.

Teachers may feel the need for stronger measurement skills, but perceive that the available formal coursework is largely irrelevant to their needs. A growing body of research supports this explanation (Dorr-Bremme, 1983; Gullickson & Ellwein, 1985; Gullickson & Hopkins, 1987; Salmon-Cox, 1981; Stiggins & Bridgeford, 1985; Stiggins & Conklin, 1988; Stiggins, Conklin, & Bridgeford, 1986). Most of these researchers encourage instructors of measurement courses to strive for congruency between formal instruction and teacher needs in measurement. In particular, Stiggins and his colleagues have argued persuasively that teachers are in need of particular training in how to effectively conduct rapid informal assessments in their instructional
decision making. Such assessment needs are only partially supported by current formal measurement coursework. If a teacher education program’s measurement curriculum can be made more relevant to teachers, and teachers' attitudes toward formal measurement training can be improved, then teachers should be more likely to seek more extensive training. One potential mechanism for changing the attitudes of current teachers is to require at least one entire measurement course at the undergraduate level, and to develop a curriculum for this course that is relevant to the needs of the classroom teacher. Change might then result through (a) an improvement in the measurement skills of the teacher population through the subsequent hiring of better trained teachers, and (b) current teachers noticing the improved skills of the new teachers and seeking such skills themselves, either through graduate or inservice training.

There are signs that a more extensive requirement of formal measurement coursework will soon be adopted by many teacher education programs. A joint committee of AACTE, AFT, NCME, and NEA representatives has recently completed the Standards for Teacher Competence in Educational Assessment of Students (AACTE/AFT/NCME/NEA Joint Committee, 1989) the development of which is described in the chapter by James Sanders in this volume (chapter 7). If adopted, the Standards may serve as the needed impetus for curricular change in teacher education programs. Such changes would have a profound impact on the measurement training of aspiring teachers, gradually leading to improvement in the skill levels of the population of practicing teachers.

REFERENCES


I began thinking about the relationship between measurement knowledge and teacher effectiveness a few years ago when our Teachers College was considering curriculum changes in our undergraduate teacher education program. Many questions about the amount and type of measurement knowledge to be included in our teacher preparation programs were raised and discussed. The recent Buros-Nebraska Symposium on Measurement and Testing related to this topic and the chapters included in this volume have resulted in further consideration of this issue. My review and analysis of this information indicates that there are many unanswered questions about the relative importance of measurement knowledge for prospective teachers. Research in other areas of education and psychology suggest that knowledge (of measurement or whatever else one chooses) may contribute only a very small percentage of the variance to that which is effective teaching.

My primary objective for this chapter is to provide an alternative perspective on how measurement training should be undertaken with teachers. Towards that end a review of what is known about effective teachers and the implications of this information for understanding the skills that must be trained will be provided. Next, a brief overview of research from the parent training literature will be examined in order to provide some examples of how the training process for teachers might be made more efficient. One example of
the use of innovative measurement procedures and technology to improve children's academic performance is reviewed. Finally, specific suggestions for future efforts in preparation of teachers for effective educational measurement are provided.

OVERVIEW: KNOWLEDGE AND TEACHER EFFECTIVENESS

Most of the contributors to this volume feel very positively about the importance of measurement knowledge for teachers. They have written about the need for insuring that prospective teachers have adequate knowledge of relevant measurement concepts and practices. Teacher effectiveness has been suggested to be related to knowledge of assessment practices (e.g., Stiggins, chapter 2, this volume), grading procedures (e.g., Terwilliger, chapter 4, this volume), classroom evaluation (e.g., Gullickson, chapter 1, this volume), and testing (e.g., Marso & Pigge, chapter 6, this volume). The issue of teacher knowledge in educational assessment is felt by some to be important enough to develop and promote standards for teacher competence in educational assessment of students. There is a widespread belief among both general educators and measurement experts that teachers are not very knowledgeable about educational measurement and there are data available in support of these beliefs (e.g., Wise & Lukin, chapter 7, this volume).

The available data and the overwhelming sentiment that teachers are deficient in measurement training give testimony to the need to do something different. The point of this chapter is not to argue against the need for change, but to suggest an alternative approach to how and what teachers need to learn about measurement. The feelings run high, but data do not appear to be available to demonstrate a clear, convincing relationship between extent of measurement training and quality of teaching and learning. What is missing from the articles referred to above and the current research literature is evidence that improvement in teachers' knowledge of measurement will result in (a) better measurement in the classroom, (b) more effective teaching, or (c) children who learn more. Perhaps these data are forthcoming, but I doubt if increasing teachers' knowledge of measurement principles is the answer to improving teachers' measurement skills in classrooms.
IN SEARCH OF EFFECTIVE TEACHERS

The past few years have seen much comment on and study of the characteristics of effective instructional environments (e.g., Bickel, 1990; Bloom, 1984; Greenwood, Delquadri, & Hall, 1984; McKee & Witt, 1990). Teacher effectiveness has been a topic of special interest to many within the education establishment (Brophy & Good, 1985; Evertson, 1987; Walberg, 1985). As a result of this research there are few among us who would argue with the assertion that teacher behavior and classroom organization influence student productivity. It has not always been so. In the past a child’s ability in the classroom was seen to be a function of their intelligence, their style of learning, their personality, and their behavior in the classroom as opposed to a function of teacher skill. For example, it has been much more common to hear people talk of child deficits in learning (e.g., mental retardation, learning disabilities, behavioral impairment, slow learner) than teacher deficits in teaching (McKee & Witt, 1990).

During the decade of the 1980s researchers and practitioners began to attend to teacher effectiveness with greater vigilance (Brophy & Good, 1985). Much has been accomplished and a clearer picture of a teacher’s contribution to learning has emerged. Similarities have been noted between the strategies that are effective in both regular and special classrooms (e.g., Bickel & Bickel, 1986; U.S. Department of Education, 1986). Two general areas of skill development that have been shown to be of central importance in teacher effectiveness are classroom management (e.g., Evertson, 1987; Gettinger, 1988) and quality of instruction (e.g., McKee & Witt, 1990; Walberg, 1985).

Establishing/Maintaining Classroom Management

The importance of a teacher having an effective system of classroom management has been demonstrated beyond a reasonable doubt. There is little question that classrooms in which children follow rules and engage in high rates of appropriate behavior are classrooms where students are “set up” to learn (Evertson, 1987; Martens & Witt, 1988). The reader should not be deceived, however, for “classroom management” is neither a single nor simple skill and involves much more than just keeping children quiet and obedient. Teachers who skillfully manage a classroom use many different skills and subtle combinations of skills in complex patterns that are only just beginning to be understood (Sharpe & Hawkins, in press). It is clear that classroom management comprises many different components that
when used effectively lead to cost-effective management of an instructional environment.

It has been shown, for example, that an effective management system involves considerable analysis by teachers of instructional activities and expected student behaviors before ever entering the classroom. Management procedures to be implemented during the first few weeks of school must be planned carefully and strategies must be developed to maintain the rules established during the first weeks of the school year (McKee & Witt, 1990). Specifically, teachers must determine and define clearly classroom rules and decide what consequences will be imposed for infractions, care must be taken to plan procedures and establish requirements for everyday routines, teachers must provide procedures to maintain student accountability, and teachers must manage both appropriate and inappropriate behavior.

As indicated earlier, teachers who demonstrate these skills have consistently been identified as more effective teachers. Most of this research has been completed during the last 20 years and the understanding of the complexity of being an effective teacher is only beginning to emerge. Much has been learned; however, analysis of the relative importance of various ecological (e.g., classroom size, building climate, class content, student background) and individual (e.g., personality, knowledge, skills) characteristics is in its infancy.

Quantity and Quality of Instruction

Not only must teachers manage the behavior of the classroom effectively, they are expected to teach students specific content. Analysis of instructional quantity and quality has been undertaken by numerous investigators during the past few years and has demonstrated that lessons that proceed smoothly, are well paced, and maintain high student engagement contribute to an effective learning environment (Greenwood et al., 1984; McKee & Witt, 1990; Walberg, 1985). A review of this research makes it clear that teachers who allocate more time for instruction have classrooms where more instruction is delivered, students who engage in high rates of academic responding tend to have the highest achievement rates, quick and frequent teacher feedback and correction is positively related to student productivity, teachers who are able to present material and instructions clearly and relatively quickly are more effective than those who cannot, and independent practice by students during free time or via homework assignments increases academic skill develop-
ment. An additional benefit of this research has been the shift from focus on student deficits in learning to teacher skill in teaching.

Summary of Teacher Effectiveness Research

The goal of this analysis of teacher effectiveness was to uncover the kinds of skills that lead to someone being identified as an effective teacher. Although the literature review provided above is not exhaustive, it does appear that most of what we know about teacher effectiveness relates to how teachers behave while in the classroom. That is, teachers who engage in certain behaviors in the presence of students tend to be more likely to produce student learning than teachers who engage in other behaviors.

The picture that emerges is that of a teacher who plans before entering the classroom and who has a clear sense of student expectations and a set of rules for classroom performance. The effective teacher is an active, engaged individual who delivers instruction clearly and demonstrates what she or he expects. The teacher moves around the class and closely monitors student performance.

Little evidence is available that relates teacher knowledge to management skill and instructional effectiveness. As we will soon see, this tentative relationship between knowledge and practice is evident in other research. In some areas of education (e.g., educational measurement) we do have evidence of the levels of teacher knowledge of basic principles, concepts, and practices (e.g., Gullickson, 1986; Schafer & Lissitz, 1987; Wise & Lukin, chapter 7, this volume), but a clear relationship between levels of knowledge and teacher behavior has not been established. That is, do teachers who know more teach better? Do teachers need to know and understand effective practices in order to implement these practices? At present there is little information that would support an affirmative answer to these questions.

TRAINING TEACHERS: LESSONS FROM PARENT TRAINING

Just as the characteristics of effective schools and classroom teachers have come under intense scrutiny, parenting and parents have been studied relentlessly during the past quarter century and this literature has been the focus of numerous reviews (e.g., Bernstein, 1983; Budd & Fabray, 1985; Kramer, 1990; Moreland, Schwebel, Beck, & Wells, 1982; O'Dell, 1985). One of the most important contributions
of this research has been the information yielded about how to train parents to train children. It would appear that this literature has much to offer in the search for functional information about how to train teachers (Kramer, 1990). Stated differently, understanding how to train parents to teach children should have some utility in understanding how to train teachers to teach children. Others have noted the similarities between the roles and responsibilities of parents and teachers (e.g., Becker, 1975).

Early researchers in parent training examined differences between dysfunctional and healthy parenting behavior by studying the contrast between clinic-referred and non-referred families. More recently, longitudinal analyses (e.g., Bank, Patterson, & Reid, 1987) have provided information about the development and characteristics of family systems. As a result of these efforts, a huge literature related to the training of parenting skills has emerged. Interestingly, much of this research has focused on teaching parents effective management skills (e.g., Dangel & Polster, 1984) and improving the quality of parent instruction (e.g., Wahler & Dumas, 1984). Management of child (student) behavior and quality of the instructional environment are the variables discussed earlier as being central to effective teaching.

Many different strategies have been evaluated in an effort to train parents to be better behavior change agents with their children. Strategies have included verbal instruction, written materials, modeling, role playing, and rehearsal, as well as innumerable combinations of these approaches. Many different authors have reviewed these findings and a better understanding of the factors that promote skill development in parents has emerged (e.g., O'Dell, 1985; Kramer, 1990).

Verbal Learning

The term “verbal” has been used to describe a group of strategies that includes discussion groups, written materials, brief lectures, or similar approaches that involve talking to or with parents and/or having them read materials (Kramer, 1990). These strategies have the advantage of being relatively easy to deliver and have been used extensively by practitioners for many years. Many parents want or need assistance, and verbal strategies allow large numbers of parents to be reached with a minimal amount of time and personnel.

Studies that have systematically compared different instructional formats have shown that verbal formats (verbal instruction, lectures,
reading, etc.) are among the least effective (Flanagan, Adams, & Forehand, 1979; O'Dell, Flynn, & Beniolo, 1977). It is interesting to note that this literature does demonstrate that these verbal approaches do help parents increase their knowledge of effective parenting and teaching techniques and that this knowledge can be imparted rather quickly. Unfortunately, the evidence also suggests that this knowledge does not routinely translate into effective behavior in the natural environment.

Reviewers have concluded that talking to parents, as is often done in individual therapy and short-term workshops, does not promote behavior change in a consistent manner. This is true even when instruction is provided by an “expert” (Ziarnik & Bernstein, 1982). This finding is clear across many different training formats and contexts (see Kramer, 1990 for a more complete review). Nor is there any evidence that having parents read published texts or self-help manuals promotes behavior change in parents or their children (McMahon & Forehand, 1980). Reading materials and verbal instruction have less effect on skill development than do most other approaches (O'Dell, 1985).

The similarities between the verbal instruction delivered to parents and that which occurs in many teacher education programs is obvious. Although there is evidence of change, the history of teacher training has been that teachers are taught primarily in college classrooms where instruction is delivered by experts via lectures and books. The parent training literature suggests that this strategy would result in teachers with an increased knowledge base, but that this knowledge is not very likely to manifest itself in classroom practice.

Demonstration and Participation

There is no shortage of research documenting the effectiveness of procedures that require the parent to observe and practice the skill to be learned. As in other instructional settings (e.g., driver education), procedures that require the client to be engaged in skill practice (i.e., driving) do better than those that require less direct involvement (i.e., reading the rules about driving). Both modeling (e.g., Nay, 1975; Webster-Stratton, 1981) and role playing/rehearsal (e.g., Flanagan et al., 1979; O'Dell, Flynn, & Beniolo, 1977) have been shown to enhance learning. In addition to the importance of having an opportunity to view a model and/or practice, the presence of corrective feedback generally enhances training effects (Bernal, Williams, Miller, & Reagor,
Homework that involves practice of specific skills has also been shown to enhance parent training effects (Forehand & McMahon, 1981).

Of course, all of the training approaches cited above involve verbal instruction. In addition, they have the characteristic of requiring the subject to practice (i.e., engaged time) the skill to be learned. "Engaged time," as we saw earlier, appears to be critical to the development of skills in children as well as parents. These strategies that involve demonstration and practice have been used to assist parents in solving a myriad of problems (see, for example, O'Dell, 1985 or Kramer, 1990).

Summary of Parent Training Research

The hope was that this analysis of the parent training literature might provide some direction in regard to the most effective strategies for training teachers. This review revealed that parents tend to become better at implementing behavior change strategies when they are trained via methods that involve demonstration and practice. Verbal methods, when used in isolation or in combination with other verbal methods, are not very effective at promoting skill development in parents. Knowledge of basic principles can be imparted rather quickly and information can be an important adjunct to the training process. As has been suggested, however, knowledge does not appear to be a very good predictor of ability to implement skills in applied settings.

USING MEASUREMENT TO MONITOR PROGRESS AND IMPROVE ACHIEVEMENT

Not only have teachers and teacher preparation programs been criticized for lack of attention to educational measurement, much dissatisfaction has been expressed with the measurement tools available to teachers and other school professionals (e.g., school psychologists) interested in assessing student progress and response to academic interventions. Although standardized test batteries, criterion-referenced instruments, and informal assessment inventories have been used to measure student achievement and to diagnose specific skill strengths and weaknesses, these tools have not been very useful for measuring short-term change in student academic responding (Lentz, 1988). Many have questioned the technical properties of
the most widely used achievement tests and whether these instru-
ments should be used for any type of educational decision making
(Fuchs, Fuchs, Benowitz, & Berringer, 1987; Ysseldyke, Algozzine,
Regan, & Potter, 1980). Furthermore, there is often a mismatch
between the content and sequence in which skills are introduced in a
particular school district and the content of achievement tests and
inventories.

In response to these difficulties, a number of approaches have
been developed that focus on direct observation and measurement of
academic skills (Becker, Engelmann, Carnine, & Maggs, 1982; Deno,
In general, these approaches have emphasized direct, repeated assess-
ment of academic target behaviors (Lentz, 1988). Recently, much
attention has been devoted to the technology of curriculum-based
measurement (e.g., Shinn, 1989; Tucker, 1985) and the use of this
technology in monitoring the development of children's basic aca-
demic skills (e.g., reading, spelling, written expression, and arith-
metic). In fact, the 1980s saw a virtual explosion of research in
curriculum-based measurement. A brief overview of this research
and examples of potential use in classrooms are provided in the
following sections.

Curriculum-Based Measurement

The term curriculum-based measurement (CBM) has been most
closely associated with research completed at the University of Min-
nesota (e.g., Deno, 1985). CBM is one of several types of
curriculum-based assessment strategies that have been utilized dur-
ing the past few years. One of the major goals of these efforts has been
to insure a match between the content of academic assessments and
the content of the local curriculum. With regard to CBM, researchers
wanted to develop a technology for assessing student achievement
that was reliable and valid, simple and efficient, easily understood,
and inexpensive.

Although it is beyond the scope of this chapter to detail the nature
of the entire CBM research process (see Shinn, 1989; or Tucker, 1985
for more detail on the research on CBM), it is clear that the develop-
ment of CBM has led to the existence of a technology where academic
probes of 1-3 minute duration can be developed from curriculum
materials, be used by teachers in a reliable manner, and provide
accurate indicators of student progress (e.g., Deno, 1985). For ex-
ample, research has shown that counting the number of words read
correctly from passages selected from a child’s basal reader during brief (1-2 minute) oral reading sessions that are repeated once or twice a week provide an excellent indication of a child’s progress in reading (Deno, 1985). In addition to reading, investigation of curriculum probes have been conducted across a variety of academic skill areas including spelling (e.g., Fuchs, Fuchs, Hamlett, & Allinder, 1991), written expression (e.g., Deno, Marston, & Mirken, 1982), and arithmetic (e.g., Fuchs, Fuchs, Hamlett, & Steckler, 1990). CBM research has been disseminated widely, with applications in special (e.g., Germann & Tindal, 1985) and regular (e.g., Marston & Magnusson, 1985) education.

The evidence is clear that CBM investigations have produced more direct and cost-efficient methods (as compared to available tests of achievement) of monitoring student progress. Indeed, the data obtained in the Minnesota investigations suggest that curriculum-based probes “are as psychometrically sound as standardized achievement tests, simpler to administer, and are much less expensive” (Lentz, 1988, p. 98). CBM measures have been applied successfully to screening for program eligibility (e.g., Marston & Magnusson, 1985), placement in curriculum levels (e.g., Deno & Mirken, 1977), and most prominently, progress monitoring (e.g., Fuchs, 1989). CBM data have been used to differentiate among exceptionalities and place children in special programs (Marston & Magnusson, 1985; Shinn & Marston, 1985). Still others have advanced methods of developing local CBM norms to assist individual school districts in the identification and placement of children in special programs (e.g., Shinn, 1988). As this chapter is being written, I am aware that development of local curriculum-based measurement normative data is occurring in at least two school districts in Nebraska and is under discussion in many others.

Until recently, little attention has been given to using CBM to assist classroom teachers in determining the effectiveness of instruction. It is this research by Fuchs and colleagues (e.g., Fuchs, 1993; Fuchs, Fuchs, & Hamlett, 1989) that appears especially promising as we search for best practices in classroom measurement and strategies for teaching teachers measurement strategies that are both efficient and effective.

**Computer-Managed/Measurement-Guided Instruction**

Although CBM has been presented as a better mousetrap, there is little information available to suggest that teachers will use it. Unfor-
tunately, education does not have a long history of adopting efficient procedures (e.g., Lentz & Kramer, 1993). Teachers who collect student performance data typically do not use these data to evaluate and alter instruction (Baldwin, 1976; White, 1974). Obviously, any attempt to use CBM to impact on instructional quality must take into account the need to make the system feasible for teacher implementation.

Lynn Fuchs and colleagues at Vanderbilt University have completed many studies related to these issues. For example, Fuchs, Hamlett, and Fuchs (1990) have developed and evaluated computer software applications of CBM technology in order “(1) to ensure standardization of the CBM monitoring, (2) to increase the feasibility of the monitoring systems, and (3) to extend the information teachers can derive from measurement” (Fuchs et al., 1990, p. 167). Due to availability of Apple II computer systems in many schools across the country the program is available currently only for these computers.

This software is designed to assist teachers in monitoring academic progress in reading, mathematics, and spelling. Although the CBM implementation strategies vary slightly across the three academic areas, the process of using the software looks something like this:

(1) In each of the three academic skill areas, teachers and students have separate disks. Following initial preparation of disks for individual students and orientation to the task, a student sits at the computer and completes a timed task ranging from 1 or 2 minutes for math to 2 1/2 minutes for reading to 3 minutes for spelling. The computer scores the responses and these data are saved to a student performance graph that is available for both teacher and student to observe.

(2) Following collection of baseline data, teachers are instructed to set performance goals for each student. Specific instructions are available for teachers to guide them through the goal-setting process. Teachers may select goals based on data collected during the development of this software (e.g., an average increase of .7 word per week) or their individual knowledge of the student. Teachers are encouraged to set ambitious goals for their students. When teachers view each student’s progress, they are able to see both the individual data points generated from the student’s performance and the student’s goal line (that is, the student’s hypothesized trend line based on the baseline data and the ultimate goal). Student graphs show data points but not the student’s goal line.
(3) During the school year it is recommended that students use the software once (for regular education students) or twice (for special education students) per week to provide data on the extent of their progress in whatever academic areas are being monitored. Both regular and special education students are able to use the software with little or no teacher monitoring.

(4) When teachers use their teacher disk to examine individual student data, they are prompted as follows: (a) Insufficient data for analysis--this may mean that not enough data are available for a decision or that the available data do not suggest any changes; (b) Uh-oh! Make a teaching change; or (c) OK! Raise the goal. The specific prompt depends on the amount of data that has been collected (e.g., Insufficient data . . . ) or the match between student performance and the student’s goal line (see Figure 1).

Obviously, my review of the Fuchs, Hamlett, and Fuchs (1990) Monitoring Basic Skills Progress program is very brief. I have not discussed specific strategies across the three academic areas nor looked at the specific decision rules that are the basis for making decisions about teaching or goal changes for individual students.

The primary purpose in presenting these data has been as a backdrop for pointing out that the research of these investigators (e.g., Fuchs & Fuchs, 1986) has shown that the simple graphing of CBM data as described above results in student achievement gains (outcomes of approximately .5 standard deviation units more than tabular presentation). Most importantly, requiring teachers to use standardized decision rules results in even better outcomes than just allowing teachers to visually inspect student performance data. When teachers are required to either change instructional strategies or raise goals based on computer prompts, student achievement increases (Fuchs et al., 1989). Teachers can and will use measurement to guide future instruction and it works!

Summary of CBM Research

There is little question that the development of CBM and other curriculum-based assessment strategies offers much to education, teachers, and students. The specific software application described above has been shown to improve student performance and to provide teachers with accurate assessment of student progress. This research makes clear the importance of making measurement part of
Figure 1. Examples of graphs of the type produced with Monitoring Basic Skills Progress.

Wilma
Math 1
Goal: 30
Pts: 3

Mary
Math 3
Goal: 30
Pts: 8

OK! Raise the goal.

Linda
Math 2
Goal: 40
Pts: 9

Uh-oh! Make a teaching change.

the instructional process. Little appears to be known about the extent of knowledge that the teacher participating in the studies described above possessed about educational measurement and/or curriculum-based assessment. It is clear, however, that teachers were quite capable of using the computer to keep accurate measures of students' progress and to guide instruction. Most importantly, the children who were under the guidance of these teachers obtained higher achievement scores than did those students not participating in the program.

GENERAL SUMMARY, CONCLUSIONS, AND IMPLICATIONS

There is no shortage of information indicating that teachers obtain little measurement training and that their knowledge base is limited in this domain (e.g., Gullickson, 1986; Schafer & Lissitz, 1987; Wise & Lukin, chapter 7, this volume). I agree with the other contributors to this volume about the need for improvement in measurement training for prospective teachers; however, my ideas about how to best address this need differs from theirs. I suspect that current conceptualizations of what teachers need to know about measurement and how we go about teaching measurement to teachers needs updating in light of current research findings. For example, most of what is known about teacher effectiveness relates to how the teacher behaves with a class of students and the general organizational climate of the classroom. Planning, precise instructions, modeling, role playing, guided practice, corrective feedback, reinforcement, and homework all appear related to the quality of results that an instructor can expect. The relationship between knowledge of effective practices and effective practice in classrooms is less clear. The claim is not being made here that knowledge is unimportant. For example, we have seen that knowledge obtained through written instructions can be effective adjuncts to other forms of training in helping parents reduce levels of inappropriate child behavior and that this knowledge can be imparted rather quickly. However, the relative contribution of teacher knowledge to the instructional process has not been established. How does teacher knowledge in measurement impact on student outcome? How is teacher knowledge of measurement related to teacher behavior?

It is suggested that teacher educators should spend more time studying how teachers actually measure behavior in the classroom than on measuring how much teachers know about measurement
concepts and practices. It would also be helpful to learn about how specific measurement practices influence student achievement as well as parent and student satisfaction with the instructional process. To the extent that knowledge of measurement principles is assessed, knowledge should be related to direct observations of both teacher and student performance.

Specific Implications For Teacher Training

Four specific suggestions occur based on the review of information provided above:

1. More emphasis should be placed on teaching measurement within context. It is not that teachers do not need to know about measurement, but rather that teachers should be taught about effective measurement practices as they practice teaching. Measurement should be moved out of the college classroom and into the field. College instructors should work in classrooms along with teachers to design effective and efficient measurement procedures and technologies. Such a process will have the dual benefit of making measurement seem more important to teachers and requiring college instructors to teach measurement in a manner that is useful to teachers.

2. Increased attention should be paid to the process of skill development. The information from the parent training literature would appear especially troublesome for those advocating increased amounts of didactic classroom instruction as a remedy for the measurement deficits of teachers. Parents have not been shown to be very adept at transferring learning from the instructional setting to the natural environment. This has been especially true when the instruction has involved verbal methods. Knowing what to do does not insure that parents will be able to implement effective strategies when they return to the home. It is suspected that similar findings will be shown with regard to educational measurement. Measurement should be taught by requiring teachers to do measurement and to make decisions about instruction based on the data obtained. College instructors should model appropriate behavior and provide teachers with feedback about the measurement process as it is ongoing in the classroom. Rehearsal, practice, modeling, and feedback are believed to be the key to improving the quality of measurement in schools.
3. Measurement training for teachers should hold teachers accountable for producing and measuring change in individual child behavior. The job of teaching is essentially that of producing change in individual child behavior. Much of what teachers are exposed to in educational measurement texts does not relate to evaluating change in individual student behavior or using measurement to guide instructional activity (e.g., changing instructional strategies when measurement of individual student behavior indicates that learning is not occurring). Much of the current measurement technology that teachers (and other school personnel) are exposed to for assessing achievement is insensitive to short-term change and/or unrelated to the curriculum in individual school districts.

4. Teacher training should focus on measurement of teacher behavior as well as measurement of child behavior. Although it has not been the primary focus of this chapter, the point has been made that the past few years have seen the beginning of a move away from focus on child deficits as a cause for failure to learn. The increased attention to teacher behavior has proven productive for understanding how best to impact on the quality of teaching and learning in schools and individual classrooms. Children who do not learn very well or very quickly must still be taught. We must continue to improve our efforts to train teachers that when students fail, teachers must examine and evaluate their own behavior in the search for more effective ways to impact on that particular child’s learning.

Final Thoughts

At the beginning of this chapter I indicated that my primary goal was to advance an alternative perspective on the process by which teachers should be taught about measurement. It is hoped that others have found my efforts to be productive and that the ideas advanced herein are useful as educators consider ways of improving the quality of teacher education and student productivity. Improved teaching of measurement skills and improved measurement practice in classrooms will do much to improve the quality of education being offered to children. There is much to be done and many things to be changed.
REFERENCES


Schiefelbusch (Eds.), *Teaching special children* (pp. 31-69). New York: McGraw-Hill.


Author Index

A

Adams, H. E., 209, 220
Aiken, L. R., 67, 92
Ainsworth-Land, G., 117, 125
Ainsworth-Land, V., 117, 125
Airasian, P., 18, 23, 146, 152, 177
Algozzine, B., 211, 223
Atkin, J. M., 107, 125

B

Balch, J., 131, 132, 177
Baldwin, V., 213, 219
Banathy, B. H., 109, 114, 117, 125
Bangert-Downs, R., 131, 132, 177
Bank, L., 208, 219
Barnes, S., 21, 23
Barrett, W., 98, 101, 125
Beattie, J., 67, 93
Beck, S., 207, 222
Beck, M. D., 2, 51, 155, 184
Becker W. C., 211, 219
Becker, W., 208, 219
Bellanca, J. A., 70, 92, 94

Beniolo, L. A., 209, 222
Bernal, M. E., 209, 219
Bernstein, G. S., 207, 219, 223
Bernstein, G. S., 209, 223
Bickel W. E., 205, 219
Bickel, D. D., 205, 219
Billeh, V. Y., 168, 177
Black, T. R., 168, 178
Bloom, B. S., 33, 40, 82, 85, 92, 110, 161, 162, 184, 219
Bohm, D., 112, 115, 116, 125
Borg, W. R., 43, 61, 136, 152, 178
Brickell, J., 143, 180
Bridgeman, B., 133, 178
Brophy, J., 129, 178, 205, 219
Bruce, F. A., 157, 180
Budd, K. S., 207, 219
Bull, K. S., 134, 183
Burke, J., 99, 108, 125
Burry, J., 151, 181
Bushman, A., 133, 178

C

Cadwell, J., 152, 183
Calfee, R., 107, 108, 125
Calhoun, M. L., 67, 93
Callendar, M. G., 111, 112, 127
Capra, F., 109, 125
Carnine, D. W., 211, 219
D'Ydewalle, G., 132, 178
Dangel, R. F., 208, 219, 223
Dawe, R., 104, 126
DeCorte, E., 132, 178
Delquadre, J. C., 205, 206, 221
Deno, S. L., 211, 212, 219, 223
Dentuch, M., 71, 93
Diamond, E. E., 140, 178
DiBernardo, G., 104, 125
Dorre-Bremme, D., 107, 126, 136, 143, 144, 146, 149, 150, 151, 152, 158, 178, 200, 201
Duchastel, P. C., 132, 183
Dumas, J. E., 208, 222
Easley, J., 7, 17, 25
Eaton, M., 211, 221
Ebel, R. L., 47, 48, 61, 74, 93
Edelsky, C., 104, 126
Elliot, E., 42, 62
Ellwein, M. 3, 134, 150, 179
Elton, L., 129, 178
Engelmann, S., 211, 219
Engler, M. D., 82, 92
Erlich, P., 105, 106, 114, 115, 127
Evertson, C. M., 205, 220
E

Fabray, P. L., 207, 219
Farnham-Diggory, S., 110, 126
Farr, R., 48, 49, 61
Flanagan S., 209, 220
Fleming, J., 6, 23, 44, 45, 46, 47, 61, 89, 93, 134, 138, 178, 149, 161, 199, 201
Floden, R., 150, 180
Flynn, J. M., 209, 222
Forehand, R., 209, 210, 220, 222
Freeman, D., 150, 180
Fremer, J., 140, 178
Friedman, S. J., 140, 179
Frisbie, D. A., 20, 25, 91, 94, 140, 179
Frisbie, D., 25, 94, 184
Fuchs, D., 214, 220
Fuchs, L. S., 212, 213, 214, 220
Furst, E. J., 82, 92

G

Gardner, H., 117, 120, 126
Gardner, J., 68, 75, 93
Gay, A., 157, 180
Gay, L. R., 132, 157, 179
Germann, G., 212, 220
Germundsen, R., 22, 24
Gettinger, M., 205, 221
Gleick, J., 113, 126
Glenn, A. D., 22, 24
Good T. L., 129, 178, 205, 219
Goslin, D. A., 1, 24, 42, 61
Green, K., 24, 145, 153
Greenwood, C. R., 205, 206, 221
Griswold, M., 6, 20, 25, 94, 168, 184
Gronlund, N. E., 6, 19, 24, 85, 86, 93, 175
Guillickson, A. R., 3, 24, 43, 44, 49, 51, 134, 136, 141, 143, 144, 150, 152, 153, 188, 200, 204, 207
Guza, D. J., 132, 179

H

Haertel, E., 7, 8, 9, 17, 18, 20, 21, 24, 148, 150
Hall, B. W., 151, 180
Hall, R. V., 205, 206, 221
Halpin, G., 132, 167, 179, 180
Hamlett, C. L., 212, 214, 220
Hampel, R. L., 110, 126
Hansen, C., 211, 221
Hargraves, A., 104, 126
Haring, N. G., 211, 221
Harman, S., 104, 126
Hawkins, A., 205, 222
Hechinger, F., 107, 126
Herman, J. L., 107, 126, 149, 180
Hermanowicz, H. J., 144, 180
Hiebert, E., 107, 108, 125
Hill, K. T., 133, 180
Hill, W. H., 82, 92
Hills, J. R., 52, 65, 93, 157, 180
Hopkins, K. D., 3, 144, 179
Houston, J., 98, 126
Hsiao, F. S., 67, 93
Hutchins, C. L., 103, 104, 126
Author Index

I
Impara, J. C., 157, 180
Infantino, R. L., 158, 180
Izu, T.

J
Jarman, B., 110, 111, 112, 122, 124, 126
Johnson, D., 71, 93, 118
Johnson, R, 71, 93, 118

K
Keating, P., 103, 104, 106, 126
Kellaghan, T., 145, 152, 180
Kelly, J. L., 1, 2, 25, 42, 47, 50, 51, 52, 61, 141, 183
Kennedy, D., 107, 125
Kinney, D., 143, 180
Kirby, P. C., 163, 183
Kirp., D. L., 105, 106, 115, 127
Kirschenbaum, H., 70, 83, 93
Klein, J. T., 67, 93
Klimko, I. P., 134, 180
Kodras, J. E., 66, 94
Kramer, J. J., 207, 208, 209, 210, 213, 221
Krathwohl, D. R., 82, 92
Kuhn, T., 97, 99, 101, 110, 114, 116, 119, 126
Kuhs, T., 150, 180
Kulik, C. C., 131, 132, 133, 155, 177, 181
Kulik, J. A., 131, 132, 133, 155, 177, 181
Kundert, D., 134, 183

L
Lambert, R. F., 142, 181
Land, G., 110, 111, 112, 122, 124, 126
Lane, D. S., 134, 183
Laurillard, D., 129, 178
Lazar-Morrison, C., 151, 181
Lehmann, I. J., 25, 42, 61, 133, 141, 148, 182
Lentz, F. E., 211, 212, 213
Liberty, K. A., 211, 222
Lieberman, A., 105, 126
Linn, R. L., 64, 93, 133, 136, 150, 179, 181
Litvak, S., 105, 110, 114, 126
Liverman, M., 157, 180
Lovitt, T., 211, 221
Lynn, M. A., 143, 180

M
Madaus, G. F., 145, 152, 180
Maggs, A., 211, 219
Magnusson, D., 212, 221
Marston, D., 212, 219, 221
Martens, B. K., 205, 221
Marzano, R. J., 102, 126
Maturana, H. R., 122, 124, 127
Mayo, S. T., 1, 24, 42, 45, 61, 141, 156, 157, 173, 182
McKee, W. T., 205, 206, 221
McLaughlin, T. F., 132, 179
Medina, N. J., 107, 127
Mehrens, W. A., 1, 2, 25, 42, 47, 50, 51, 52, 61, 133, 141, 148, 182, 183
Miller, W. H., 209, 219
Mirken, P. K., 212, 219
Mitzel, H. E., 64, 93
Monk, J. J., 132, 134, 182
Moreland, J. R., 207, 222
Moy, R., 151, 181
Moynihan, P., 69, 93

N
Napier, R., 70, 83, 93
Nash, W., 133, 178
National Governors' Association, 106, 127
Nay, W. R., 209, 222
Neill, D. M., 107, 127
Newman, D. L., 134, 183
Newman, D. C., 136, 149, 156, 157, 182, 199, 201
Nickerson, R. S., 134, 183
Nitko, A., 21, 24, 52
Noll, V., 9, 25, 42, 61, 188, 189, 202
Nungester, R. J., 132, 183

O
O'Dell, S. L., 207, 208, 209, 210, 222
Oakes, J., 103, 104, 106, 126
Oescher, J., 163, 183
<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oosterhof, A.C.</td>
<td>90, 93</td>
</tr>
<tr>
<td>Ornstein, R.</td>
<td>105, 106, 114, 115, 127</td>
</tr>
<tr>
<td>Patrick, C.L.</td>
<td>107, 125</td>
</tr>
<tr>
<td>Patterson, G.R.</td>
<td>208, 219</td>
</tr>
<tr>
<td>Peat, F.D.</td>
<td>112, 115, 116, 125</td>
</tr>
<tr>
<td>Peckham, P.D.</td>
<td>132, 183</td>
</tr>
<tr>
<td>Pigge, F.L.</td>
<td>129, 134, 136, 138, 141, 142, 143, 145, 146, 149, 152, 159, 161, 204</td>
</tr>
<tr>
<td>Polin, L.</td>
<td>151, 181</td>
</tr>
<tr>
<td>Porter, A.</td>
<td>150, 180</td>
</tr>
<tr>
<td>Porter, A.C.</td>
<td>1, 2, 25, 42, 47, 50, 51, 52, 61, 141, 183</td>
</tr>
<tr>
<td>Potter, M.</td>
<td>211, 223</td>
</tr>
<tr>
<td>Prather, J.E.</td>
<td>66, 94</td>
</tr>
<tr>
<td>Presseisen, B.Z.</td>
<td>82, 86, 94</td>
</tr>
<tr>
<td>Quellmalz, E.S.</td>
<td>33, 40</td>
</tr>
<tr>
<td>Reagor, P.A.</td>
<td>209, 219</td>
</tr>
<tr>
<td>Regan, R.</td>
<td>211, 223</td>
</tr>
<tr>
<td>Reid, J.B.</td>
<td>208, 219</td>
</tr>
<tr>
<td>Roe, M.D.</td>
<td>132, 183</td>
</tr>
<tr>
<td>Roeder, H.H.</td>
<td>42, 49, 61, 144, 183, 202</td>
</tr>
<tr>
<td>Rogers, E.M.</td>
<td>129, 183</td>
</tr>
<tr>
<td>Rosenshine, B.</td>
<td>129, 183, 220</td>
</tr>
<tr>
<td>Ruddell, R.B.</td>
<td>141, 183</td>
</tr>
<tr>
<td>Rudman, H.C.</td>
<td>1, 2, 25, 42, 47, 50, 51, 52, 61, 141, 183</td>
</tr>
<tr>
<td>Russell, P.</td>
<td>105, 111, 112, 116, 127</td>
</tr>
<tr>
<td>Saigh, P.</td>
<td>51, 133, 183</td>
</tr>
<tr>
<td>Salmon-Cox, L.</td>
<td>136, 143, 149, 183, 200, 202</td>
</tr>
<tr>
<td>Sanders, J.R.</td>
<td>41, 42, 52, 62, 201</td>
</tr>
<tr>
<td>Sax, G.</td>
<td>132, 183</td>
</tr>
<tr>
<td>Schafer, W.D.</td>
<td>9, 25, 43, 49, 62, 144, 145, 183, 188, 192, 194, 202, 207, 222</td>
</tr>
<tr>
<td>Schmidt, W.</td>
<td>150, 180</td>
</tr>
<tr>
<td>Schofield, W.</td>
<td>70, 94</td>
</tr>
<tr>
<td>Schwab, A.I.</td>
<td>207, 222</td>
</tr>
<tr>
<td>Schwille, J.</td>
<td>150, 180</td>
</tr>
<tr>
<td>Scriven, M.</td>
<td>63, 67, 94</td>
</tr>
<tr>
<td>Semzee, A.W.</td>
<td>105, 110, 114, 126</td>
</tr>
<tr>
<td>Sharpe, T.L.</td>
<td>205, 222</td>
</tr>
<tr>
<td>Shavelson, R.J.</td>
<td>152, 183</td>
</tr>
<tr>
<td>Sheldon, H.W.</td>
<td>143, 184</td>
</tr>
<tr>
<td>Shinn, M.R.</td>
<td>211, 212, 222</td>
</tr>
<tr>
<td>Shulman, L.</td>
<td>150, 184</td>
</tr>
<tr>
<td>Simon, S.B.</td>
<td>70, 94</td>
</tr>
<tr>
<td>Simon, S.</td>
<td>70, 83, 93</td>
</tr>
<tr>
<td>Sizer, T.</td>
<td>106, 118, 127</td>
</tr>
<tr>
<td>Slavin, R.</td>
<td>71, 94, 104, 127</td>
</tr>
<tr>
<td>Smith, G.</td>
<td>66, 94</td>
</tr>
<tr>
<td>Snyder, B.R.</td>
<td>131-132-183</td>
</tr>
<tr>
<td>Spady, W.G.</td>
<td>71, 73, 94</td>
</tr>
<tr>
<td>Sproull, L.</td>
<td>142, 184</td>
</tr>
<tr>
<td>Stager, S.</td>
<td>24, 145, 153</td>
</tr>
<tr>
<td>Stake, R.</td>
<td>7, 17, 25</td>
</tr>
<tr>
<td>Stallings, W.M.</td>
<td>132, 134, 136, 149, 156, 157, 182, 199, 201</td>
</tr>
<tr>
<td>Starch, W.D.</td>
<td>42, 62</td>
</tr>
<tr>
<td>Stecker, P.M.</td>
<td>212, 220</td>
</tr>
<tr>
<td>Stetz, F.P.</td>
<td>2, 50, 51, 155, 184</td>
</tr>
<tr>
<td>Stiggins, R.</td>
<td>6, 10, 18, 19, 20, 21, 25, 27, 28, 40, 43, 51, 62, 91, 130, 140, 142, 144, 146, 149, 151, 152, 153, 168, 173, 184, 199, 200, 202, 204</td>
</tr>
<tr>
<td>Stiles, D.</td>
<td>104, 125</td>
</tr>
<tr>
<td>Swerts, A.</td>
<td>132, 178</td>
</tr>
<tr>
<td>Szafran, R.</td>
<td>133, 184</td>
</tr>
<tr>
<td>Takeuchi, R.T.</td>
<td>158, 184</td>
</tr>
<tr>
<td>Terwilliger, J.S.</td>
<td>63, 66, 81, 87, 90, 94, 204</td>
</tr>
<tr>
<td>Thorndike, R.L.</td>
<td>73, 87, 94, 110, 178</td>
</tr>
<tr>
<td>Thornton, C.L.</td>
<td>133, 184</td>
</tr>
<tr>
<td>Timar, T.B.</td>
<td>105, 106, 115, 127</td>
</tr>
<tr>
<td>Tindal, G.</td>
<td>212, 220</td>
</tr>
<tr>
<td>Titus, A.</td>
<td>133, 178</td>
</tr>
<tr>
<td>Travis, J.W.</td>
<td>111, 112, 127</td>
</tr>
<tr>
<td>Trentham, 133, 184</td>
<td></td>
</tr>
<tr>
<td>Tucker, J.A.</td>
<td>211, 222</td>
</tr>
<tr>
<td>Tyler, R.W.</td>
<td>143, 184</td>
</tr>
<tr>
<td>U.S. Department of Education</td>
<td>205, 222</td>
</tr>
<tr>
<td>Valcarce, R.W.</td>
<td>43, 61, 136, 152, 178</td>
</tr>
<tr>
<td>Varela, F.J.</td>
<td>122, 124, 127</td>
</tr>
<tr>
<td>Veenman, S.</td>
<td>152, 184</td>
</tr>
</tbody>
</table>
W

Wahler, R. G., 208, 222
Walberg, H. J., 205, 206, 222
Wanous, D. S., 1, 2, 25, 42, 47, 50, 51, 52, 61, 141, 183
Wanous, D. S., 25, 42, 61, 141, 184
Warren, J. R., 88, 91, 94
Webster-Stratton, C., 209, 222
Wells, K. C., 207, 222
Wexley, K. N., 133, 184
White, O. R., 211, 222
Wigfield, A., 133, 180
Wikelund, K., 6, 20, 25, 94, 168, 184
Williams, D. E., 209, 219
Witt, J. C., 205, 206, 221
Wittrock, M. C., 64, 95, 178, 223
Wolmut, P., 188, 202
Worthen, B. R., 43, 61, 136, 152, 178
Wrinkle, W. L., 75, 95

Y

Yeh, J. P., 150, 185
Ysseldyke, J. E., 211, 223

Z

Ziarnik, J. P., 209, 223
Zubrow, D., 142, 184
Subject Index

A

achievement targets in assessment terms., 32
American Association of Colleges for Teacher Education (AACTE), 42, 49, 135, 188, 201
American Federation of Teachers (AFT), 42, 135, 201
assessment, 41
achievement targets, 32
to be assessed, 33
classroom uses of, 32
competencies, 31
feedback, 200
inservice training, 195, 198
predictive uses of, 106
task demands of classroom, 28, 31
the qualities of, 34
assessment competence standards of, 42
assessment literacy
specific actions to promote, 38-40
assessment results, 32, 54, 57-58, 135
communicating of, 58
feedback on, 37
school improvement, 57
assessment tools
observations of professional judgements
performance assessments, 35
paper-and-pencil assessment instruments, 33, 35, 130
performance assessments, 56, 56-57, 151
personal communication, 35
assessment training, 28
associated risks, 30
certification regulations, 28-29
rationale for teachers, 38
reasons for neglect, 29
quality of assessments, 30
assessments
student, 22, 41-42, 58, 61
assigning grades, 64-67, 78, 83, 90, 92, 136, 149

B

behavioral objectives, 110
beliefs about testing, 152-155

C

certification regulations
for assessment training, 28
classroom assessment(s), 27-37,
39-40, 46, 53, 55, 65, 107, 135, 151,
classroom teachers
standards for, 42, 135
classroom uses of assessment, 32
classroom-based evaluation, 1, 19
I

instruction, 42, 46-52, 54, 56-57, 60, 64, 68, 77, 81, 83, 85-86, 100, 104-105, 117, 122-123, 129, 136, 143-146, 149-150, 153, 158, 173, 175, 189-192, 199, 200, 205-217

instructional environments, 205

interpersonal dimensions of classroom assessment, 35-36

in-service training, 143, 146, 158, 170, 172

item-writing faults, 157

K

kinds of achievement targets to be assessed, 33

L

letter grades, 75, 85, 149-150, 172

training, 203

skills, 203

measurement and evaluation

formal training in, 144

measurement competencies needed by classroom teachers, 47, 138, 156

measurement instruction, 1, 3, 8-9, 11, 14, 19, 23, 45, 51, 198

mechanistic-reductionist schools, 100-101, 105, 107, 108

curriculum, 100-101

effective teaching, 103

instructional practices, 103

teachers purpose in assessing, 102

minimal vs. developmental objectives, 85-87

teachers purpose in assessing, 102

minimal vs. developmental objectives, 85-87

teachers purpose in assessing, 102

minimal vs. developmental objectives, 85-87

M

mathematics anxiety, 200

measurement

knowledge of, 49, 158, 176, 187, 193-195, 203
feedback
assessment results, 32

N
National Council on Measurement in Education (NCME, 41
National Council on Measurement in Education (NCME), 140
National Education Association (NEA), 42, 135, 201
National Teacher Examination (NTE), 29
nature of reality, 99
change, 109

O
observations of and professional judgments, 35
outcome-based education, 71-73
overreliance on test scores, 152

P
paradigm shift, 110-111, 114
parenting, 207
parents, 207
parent training, 203, 207-210, 217
strategies, 208
placement, 1, 3, 6, 8-9, 11, 14-15, 17-23, 27,
30-32, 38-42, 47, 134, 212
preparation of teachers, 2, 3, 9, 42, 53,
107, 120, 156, 204
rationale for, 38

Q
question structure quality, 130
question writing skills, 142
question types, 150, 157, 159-171

S
school improvement, 57
scientific view, 109
scientific method, 109
specialized courses, 79
specific actions
to promote assessment literacy, 39
standardized testing, 200
standardized tests, 107, 189-192, 200
textbooks, 109
standards for classroom teachers, 42
standards of assessment competence, 42
student assessments, 41
student progress, 210
students as contributors to classroom assessment, 36
study efforts
standards for classroom teachers, 42
standards of assessment competence, 42
student assessments, 41
students as contributors to classroom assessment, 36
systematic assessment of outcomes associated risks, 30

T
task demands of classroom assessment, 31
teacher training, 27-32, 49, 51-52, 143,
146, 156-157, 161, 168, 171-175, 177,
188, 209
in assessment, 27
reasons for neglect, 29-31
testing, 42, 43
thinking skills, 86
traditional grading, 68
teacher effectiveness, 203-205, 207, 216
teacher preparation programs

course topics
attitudes, 190,
criterion-referenced, 187, 190
item analysis, 189
measurement theory, 191
norm-referenced, 187, 190
objectives, 190-191
reliability, 189-192
standardized tests, 189-192, 196
standardized test norms, 189
statistics, 189-191
teacher made tests, 191-192
test construction, 189-192
validity, 189-192
teacher training, 27-32, 49, 51-52, 143, 146, 156-157, 161, 168, 171-175, 177, 188, 209
in assessment, 27
teachers’ attitude toward testing, 143
teachers’ testing knowledge, 141
test
  cognitive demands, 134
  construction skills, 136, 156
  feedback, 133
  question arrangement, 133
  support services, 142-143, 146, 172, 176
  question difficulty, 133-134
test administration mechanics, 132
test question-writing guidelines
test specification tables, 131, 134, 153, 172
testing, 1-5, 8, 10, 13, 15, 17-18, 31, 38, 40, 42-45, 49-50, 60, 66, 85, 87-88, 101, 107, 110, 129-137, 140-146, 149-153, 156-163, 168-176, 188, 190-195, 200, 204
  competencies, 129, 136-138, 141, 159
  frequency of, 131-132
  impact of, 129, 131, 171
  thinking skills, 86, 134, 143
traditional grading, 68-74, 78, 83, 90-91
critics of, 70-74

U

University of Minnesota, 211

W

world-view, 98, 114, 116, 120, 122