6. Teachers' Testing Knowledge, Skills, and Practices

Ronald N. Marso  
*Bowling Green State University*

Fred L. Pigge  
*Bowling Green State University*

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

*Teacher Training in Measurement and Assessment Skills*. 8.  
[https://digitalcommons.unl.edu/burosteachertraining/8](https://digitalcommons.unl.edu/burosteachertraining/8)

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.
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?
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 nor 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,
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>Mean*</td>
<td>Rank</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, scoratable, 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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20.</td>
<td>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.</td>
<td>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.</td>
<td>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.</td>
<td>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.</td>
<td>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.</td>
<td>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.</td>
<td>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.
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
TEACHERS' TESTING PRACTICES

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.....)
6. TEACHERS' TESTING PRACTICES

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, day to day communication, 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...
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*</td>
<td>Rank</td>
</tr>
<tr>
<td>1. Writing good multiple-choice questions</td>
<td>3.64</td>
<td>19</td>
</tr>
<tr>
<td>2. Writing good completion questions</td>
<td>3.72</td>
<td>14.5</td>
</tr>
<tr>
<td>3. Writing good matching questions</td>
<td>3.81</td>
<td>9</td>
</tr>
<tr>
<td>4. Writing good true-false questions</td>
<td>3.58</td>
<td>20</td>
</tr>
<tr>
<td>5. Writing good essay questions</td>
<td>3.37</td>
<td>22</td>
</tr>
<tr>
<td>6. Scoring essay questions</td>
<td>3.21</td>
<td>24</td>
</tr>
<tr>
<td>7. Identifying good and poor questions for future tests</td>
<td>3.79</td>
<td>10</td>
</tr>
<tr>
<td>8. Writing questions in harmony with school and class goals</td>
<td>3.78</td>
<td>11</td>
</tr>
<tr>
<td>9. Stating objectives sufficiently clear to suggest test items</td>
<td>3.69</td>
<td>16</td>
</tr>
<tr>
<td>10. Writing test questions that demand higher thinking processes</td>
<td>3.52</td>
<td>21</td>
</tr>
<tr>
<td>11. Constructing tests that represent true student progress</td>
<td>3.65</td>
<td>18</td>
</tr>
<tr>
<td>12. Use of less formal assessments: checklists, ratings, etc.</td>
<td>3.28</td>
<td>23</td>
</tr>
<tr>
<td>13. Use of observations (visual) to assess and guide learning</td>
<td>3.72</td>
<td>14.5</td>
</tr>
<tr>
<td>14. Use of sociometric, guess who, and related techniques</td>
<td>2.88</td>
<td>26</td>
</tr>
<tr>
<td>15. Selecting good test questions from teacher manuals</td>
<td>3.93</td>
<td>5</td>
</tr>
<tr>
<td>16. Setting up readable, scorable, and attractive tests</td>
<td>3.88</td>
<td>7.5</td>
</tr>
<tr>
<td>17. Making tests reflect what is covered in text and class</td>
<td>4.23</td>
<td>3</td>
</tr>
<tr>
<td>18. Calculation of means, standard deviations, reliability, etc.</td>
<td>3.02</td>
<td>25</td>
</tr>
<tr>
<td>19. Interpreting test scores and student progress</td>
<td>3.75</td>
<td>13</td>
</tr>
<tr>
<td>20. Identifying individual and class strengths and weaknesses</td>
<td>3.91</td>
<td>6</td>
</tr>
<tr>
<td>21. Determining what needs to be retaught after tests</td>
<td>3.88</td>
<td>7.5</td>
</tr>
<tr>
<td>22. Use of tests and grades to positively influence learning</td>
<td>3.68</td>
<td>17</td>
</tr>
<tr>
<td>23. Calculating end of term grades from term work</td>
<td>4.25</td>
<td>2</td>
</tr>
<tr>
<td>24. Grading tests, papers, projects, homework, etc.</td>
<td>4.32</td>
<td>1</td>
</tr>
<tr>
<td>25. Deciding importance of tests, papers, etc. in grading</td>
<td>4.04</td>
<td>4</td>
</tr>
<tr>
<td>26. Deriving information from tests to guide students</td>
<td>3.97</td>
<td>12</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>—</td>
<td>—</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>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Short Response</td>
<td>1093</td>
<td>830</td>
<td>235</td>
<td>9</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>1</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>0</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% 11% 15% 1% .001% .001%

*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></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>496</td>
<td>100</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>100</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>100</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>
</tbody>
</table>

26 100

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>
</tbody>
</table>

106 100

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>
</tbody>
</table>

71 100

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>
</tbody>
</table>

6 100

(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>3</td>
</tr>
</tbody>
</table>

61 100

i. Test Format

<table>
<thead>
<tr>
<th>Construction Error</th>
<th>f</th>
<th>%*</th>
<th>p^1</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>
</tbody>
</table>

281 100

(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^1 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 Wiklund (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-
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 in-service 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-
develop 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


