
Arlen R. Gullickson

University of South Dakota

Follow this and additional works at: http://digitalcommons.unl.edu/burosteachertraining
Matching Measurement Instruction to Classroom-Based Evaluation: Perceived Discrepancies, Needs, and Challenges

Arlen R. Gullickson

University of South Dakota

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

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

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

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

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

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

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

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

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

TEACHER ATTITUDES AND PRACTICES

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

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

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

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

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

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

Table 1. The Median Times in Minutes Teachers Report Giving to Specified Testing Tasks for Teacher Prepared Objective Tests.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

MEASUREMENT INSTRUCTION AT THE UNDERGRADUATE LEVEL

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

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

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

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

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

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

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

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

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

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

Table 4. Means and Standard Deviations for Teachers and Professors

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


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

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

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

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

<table>
<thead>
<tr>
<th>Standardized Test Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection and use of standardized and publisher prepared tests</td>
</tr>
<tr>
<td>Norm-referenced vs. criterion-referenced tests</td>
</tr>
<tr>
<td>Test norms and interpretation based upon norms</td>
</tr>
<tr>
<td>Evaluating tests in terms of reliability and validity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nontest Evaluation Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class discussion</td>
</tr>
<tr>
<td>Observing working habits</td>
</tr>
<tr>
<td>Interpersonal relationships</td>
</tr>
<tr>
<td>Employing other evaluative devices</td>
</tr>
</tbody>
</table>

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

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

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

NEED FOR CHANGE

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

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

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

Even without changes in program structure, it seems likely that significant changes can occur. Professors appear to have considerable freedom in determining course content. Thus, if persuaded, professors
could take significant individual steps to improve undergraduate teachers' preparation in measurement and evaluation expertise. For those reasons, the needs and options for change that follow are presented only in the context of changes in the measurement and evaluation course itself.

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

OPTIONS FOR CHANGE

Tests

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

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

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

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

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

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

Teacher-made tests. Both teachers and professors appear to be in such good agreement here that it seems apparent this type of test should receive top billing in the undergraduate course. There are, however, a number of concerns that reside just under the surface. Teachers and professors profess that tests are good for all instructional decisions, formative as well as summative. Whether tests actually function to serve both formative and summative needs is open to question. Students, for example, view tests as serving
1. MATCHING MEASUREMENT INSTRUCTION

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

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

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

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

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

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

Nontest Evaluation

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

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

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

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

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

Evaluation Design

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

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

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

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

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

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

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

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

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

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

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

Technology

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

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

Student Feedback

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

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

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

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

CONCLUSION

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

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

REFERENCES

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


