Assessment Competencies of Teachers: A National Survey

Barbara S. Plake and James C. Impara, University of Nebraska–Lincoln
Jennifer J. Fager, South Dakota State University

What are the "Standards for Teacher Competence in Educational Assessment of Students"? In what domains do teachers have the greatest knowledge of measurement? Where are they least strong? Do teachers with measurement training show greater knowledge than teachers without such training?

Teachers spend considerable amounts of their time involved in assessment-related activities. In fact, Stiggins (1991) estimates that teachers may spend up to 50% of their instructional time in such activities. Recent studies relate that teachers report that they feel inadequately trained for the types of assessment demands encountered in the classroom (Wise, Lukin, & Roos, 1991). One reason for the teachers' lack of comfort engaging in assessment-related tasks could be inadequate preparation in assessment in preservice teacher education programs (Schafer & Lissitz, 1987; Schafer, 1993).

In response to concerns about the educational assessment competencies of teachers, the National Council on Measurement in Education joined forces with the National Educational Association and the American Federation of Teachers to prepare and publish Standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, NEA, 1990). These standards address seven broad skill areas in educational assessment: (a) choosing assessment methods appropriate for instructional decisions; (b) developing assessment methods appropriate for instructional decisions; (c) administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods; (d) using assessment results when making decisions about individual students, planning teaching, developing curriculum, and improving schools; (e) developing valid pupil grading procedures; (f) communicating assessment results to students, parents, other lay audiences, and other educators; and (g) recognizing unethical, illegal, and other inappropriate methods and uses of assessment information.

The W. K. Kellogg Foundation has provided NCME a grant to continue the efforts toward improving teacher competencies in the educational assessment of students. The purposes of the grant are to (a) measure teachers' competency levels in the assessment areas addressed in the Standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, NEA, 1990); (b) identify a target assessment area based on the results of measuring teachers' competencies; and (c) develop a training prototype directed at the identified target assessment area.

This article describes the activities undertaken toward meeting the first goal of the grant project: measuring teacher competency levels in the educational assessment dimensions specified in the Standards. First, instrumentation and sampling methods are described, followed by a detailed presentation of the administration procedures. Next, the results of the study are presented. Following a discussion of the results, information is shared on the next stages of the grant project. Finally, implications of the project for NCME and assessment-related in-service programs for teachers are discussed.

Procedures

Instrumentation

A two-part assessment device was developed. Part One contained items designed to assess teachers' knowledge in the competency areas identified in the Standards. Part Two asked questions relating to teachers' background and perceptions about test use.

Part One. This segment consisted of a 35-item instrument developed to measure the seven competency areas articulated in the Standards. The table of specifications was determined by identifying the knowledge, skills, and abilities specified for each of the competency areas in the Standards. One goal of item development was to prepare application questions that were realistic and meaningful to teachers' assessment.

Barbara S. Plake is the W. C. Meierhenry Distinguished Professor and Director of the Buros Institute for Mental Measurements at the University of Nebraska-Lincoln, Lincoln, NE 68512. Her specialization is applied measurement.

James C. Impara is a Professional Associate at the Buros Institute for Mental Measurements, 135 Bancroft Hall, University of Nebraska-Lincoln, Lincoln, NE 68512. His specializations are applied measurement and project evaluation.

Jennifer J. Fager is an Assistant Professor in the College of Education and Counseling at Box 507, Wenona Hall, South Dakota State University, Brookings, SD 57007-0095. Her specialization is undergraduate teacher education.
practices. Further, the administration time was to be restricted to less than 1 hour. Toward this end, initially, a 21-item instrument (with three items written for each of the competency areas of the Standards) was developed and pilot-tested on approximately 900 educational professionals in the state of Virginia.

In addition to the pilot study, two panels, each consisting of 10 NCME members, participated in a content validity study. These NCME members were selected based on the following criteria: (a) balance of professional affiliation, so that the panels consisted of academic and practitioner perspectives, and (b) prior experience with the Standards or issues related to teacher assessment literacy. The 10 members of Panel A were asked identify the competency area in the Standards best measured by each item. The 10 members of Panel B rated how well each item measured the competency area the item was designed to assess. In addition, both panels’ members were asked to rate, on a 1–6 scale, how essential the elements of the table of specifications were for teachers. Response by the panels gave confirming evidence that the items in the instrument matched well the competence-standard areas they were designed to assess.

Based on the results of the content validity study and the pilot study, several items were rewritten and additional items prepared to yield an instrument that had 5 items per standard. To better cover the multidimensional aspects of the Standards, additional items were deemed necessary. The content of these additional items was determined by the panel members’ ratings of the elements in the table of specifications. To simplify the task for the teachers, it was decided to organize the items in the test by Standard. It was felt that the teachers would be more inclined to complete the instrument if they were informed about the item’s alignment with the Standards.

The 35-item instrument was subjected to reviews by members of the project’s National Advisory Committee, practicing teachers, and advanced graduate students in a classroom-assessment class. In addition, these items were field tested on approximately 70 teachers in Lincoln, Nebraska, in fall, 1991.

**Administration.** Support for the project was obtained from the Assessment Task Force subcommittee of the Education Information and Assessment Committee (EIAC), a consortium of state agency personnel who monitor surveys and other data collection efforts by the federal government whose members also reviewed the 35-item instrument. The sampling plan involved identification of six school districts per state, selected to represent urban and rural school districts. Four of the six districts served as the primary sampling unit; the remaining two districts per state were used as alternates. Through contact with the state assessment directors, direct contact information was obtained for the district testing coordinator (or equivalent) in each sampled school district. These district testing coordinators were then requested to cooperate by randomly sampling 12 teachers from across elementary, middle, and secondary levels and six administrators from elementary and secondary school settings. A more detailed description of the sampling and administration can be found in Impara, Plake, and Fager (1993a). Teachers were given a packet of materials containing a letter requesting their participation, the instrument, an answer sheet, and a stamped, addressed envelope to use to return their answer sheets directly to the researchers.

**Sample.** From the 200 possible school systems, 98 systems from 45 states agreed to participate. The instrument was administered to a national sample of teachers and administrators in fall, 1991. A total of 555 (48%) teachers and 286 (51%) administrators returned their test materials. The distribution of respondents was widespread across the U.S., representing large and small school districts in rural, suburban, and urban areas. Returns were received from all states except Arkansas, Florida, Hawaii, Idaho, Mississippi, Missouri, and New Hampshire.

**Results**

Results are reported separately for each of the two parts of the instrument. These results are then followed by an analysis that considers differential test performance of teachers on Part One (achievement on competency areas) as a function of their responses to items from Part Two (teacher background and perceptions).

**Part One results.** Overall test performance is shown in Table 1. The internal consistency reliability estimate (KR-20) for the full test was .54. Given the criterion-referenced focus of the instrument, this value is not surprising. The purpose of the study was to report overall group performance information, rather than to provide individual assessment results for teachers. Subscales, consisting of merely 5 items each, also show low internal consistency reliabilities. Therefore, interpretations based on subscales should be considered with caution.

In general, teachers performed best in the competency area of administering, scoring, and interpreting test results; poorest overall test performance was in the area of communicating test results. Although it is recognized that individual item performance values are unreliable, some interesting trends emerged at the item level. On 10 of the 35 items, 90% or more of the teachers answered the item correctly. These items document possible areas of strength for teachers. Two of these items address issues in making assessment selection decisions. Other strength areas are found in teacher knowledge of acceptable test-taking behavior for timed test administration of a standardized test, explanation of the basis for a child’s grade to a parent, and recognition of unethical practices in standardized test administration.

On the other hand, there are a few items that were very difficult for the teachers. On five items less than 30% of the teachers answered correctly. Two of these items came from the “using assessments in grading” standard. Another difficult item focused on steps likely to increase reliability of test scores; only 13% of the teachers answered this item correctly. Most of the teachers chose an option that might arguably enhance the test’s validity (adding an essay component), but which probably was not the choice for augment-
Table 1
Test Performance

<table>
<thead>
<tr>
<th>Subscales/Standard</th>
<th>n</th>
<th>X̄</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choosing</td>
<td>555</td>
<td>3.46</td>
<td>0.93</td>
</tr>
<tr>
<td>Developing</td>
<td>555</td>
<td>3.22</td>
<td>0.80</td>
</tr>
<tr>
<td>Administering</td>
<td>555</td>
<td>3.96</td>
<td>0.90</td>
</tr>
<tr>
<td>Using-decision</td>
<td>555</td>
<td>3.40</td>
<td>1.11</td>
</tr>
<tr>
<td>Using-grading</td>
<td>555</td>
<td>3.19</td>
<td>0.78</td>
</tr>
<tr>
<td>Communicating</td>
<td>555</td>
<td>2.70</td>
<td>1.21</td>
</tr>
<tr>
<td>Recognizing ethics</td>
<td>555</td>
<td>3.26</td>
<td>0.78</td>
</tr>
<tr>
<td>Total score</td>
<td>555</td>
<td>23.20</td>
<td>3.33</td>
</tr>
</tbody>
</table>

ing reliability (increasing test length). Two other items that less than 30% of the teachers answered correctly are from the “recognizing unethical practices” standard.

Part Two results. The large majority of teachers had been in the classroom for 6–12 years. Relatively few teachers had less than 5 years of experience or more than 15. This distribution of teacher years in service is consistent with that reported by NEA and AFT (M. Berger & J. Schneider, personal communications, February 5, 1992) and therefore adds support to the representativeness of the sample of teachers in the study.

Teachers were asked to respond to some questions regarding their perception of test utility in making instructional decisions and their comfort in interpreting standardized test scores. Although 86% of the teachers tended to agree that teacher-developed tests should be used extensively to enhance instruction, only 34% felt that standardized test information should be used for the same purpose. The majority (53%) of the teachers felt at least moderately comfortable in interpreting information from standardized tests.

Nearly 70% of the teachers reported that they had taken a class, either in-service or in college, that had tests and measurement as the major focus; however, most of these teachers (54%) reported that the training was more than 6 years ago. When asked what they considered to be the best way to become more proficient in interpreting test scores, nearly 60% chose an in-service delivery option. Nearly 85% of the teachers indicated that they were at least somewhat interested in becoming more proficient in interpreting test scores and in student assessment in general.

Test Score Performance by Teacher Background and Perceptions

A series of analyses of variance (ANOVAs) was performed to investigate the possibility of differential test score performance by teachers as a function of their responses to background and perception questions in Part Two. No significant differences in overall test performance were found for teachers' perceptions of the instructional utility of teacher or standardized test information. However, a significant effect was found for teacher comfort in interpreting standardized test scores $[F(3, 547) = 3.54, p < .02]$. Teachers who tended to disagree with the statement of comfort when interpreting information from standardized tests had significantly lower overall test scores than did teachers who agreed with the statement ($X_{agree} = 22.34$; $X_{disagree} = 23.91$). The test mean for teachers who had taken a measurement class was significantly higher than that for teachers who had not ($X_{had} = 23.41$; $X_{hadnat} = 22.72$). Even though these performance values were significantly different, the actual performance difference was less than 2 points. None of the other analyses of test performance showed significantly different score results based on teacher background or perception variables.

Conclusions

The purpose of this study was to assess teachers' competencies on the seven basic assessment areas identified in the AFT, NCME, NEA (1990) Standards for Teacher Competence in Educational Assessment of Students. Overall, teachers averaged 23.20 items correct; highest performance was on the subscale measuring teacher knowledge in the areas of administering, scoring, and interpreting test results; poorest performance was on items measuring the teachers' knowledge about communicating test results. Teachers who had some measurement training scored significantly higher than those who had not; teachers expressing comfort in interpreting standardized tests scored significantly higher on the test than did teachers who expressed discomfort.

Results of this study have been used in making decisions about the next steps in the grant project. The content area of communicating and interpreting test results has been identified by the project's National Advisory Committee as the one for initial training material development. Based on the teachers' reported preference for in-service delivery, the training prototype developed will be administered under an in-service/professional development mode.

This study served an important role in identifying teacher competencies in the assessment components identified in the Standards for Teacher Competence in Educational Assessment of Students. Further analysis of teacher needs and preferences in the area of measurement could provide more data-based information for planning and delivering training programs for enhancing teachers' level of assessment competency.

Notes

This study was funded in part by a grant to NCME from the W. K. Kellogg Foundation and supported by the Buros Institute of Mental Measurements. The opinions expressed in this document do not necessarily represent those of these funding sources.

1A copy of the instrument can be obtained from the first author.

2Members of the project's National Advisory Committee are Jeffrey

Continued on page 39
D. Assisting State and Local Educators to Set Sound Assessment Policies
Presenter: Richard J. Stiggins, Assessment Training Institute

Monday, April 4, 9:00 a.m.–5:00 p.m.  Fee: $50

This training session prepares participants to be able to offer technical assistance to state and local policymakers in a manner that is likely to promote sound assessment policy and practice. The intended audience is not the policymakers themselves, but assessment specialists who are in a position to advise policymakers. As a result of participating in this training session, participants will (a) understand the basic ingredients of a sound educational policy and sound policies that are likely to promote sound assessment policy and practice. The intended audience is not the policymakers themselves, but assessment specialists who are in a position to advise policymakers. As a result of participating in this training session, participants will (a) understand the basic ingredients of a sound educational policy and sound regulations that support policy implementation, (b) know and be able to share the basic elements of an effective school district and/or state-level assessment policy, (c) understand the differences in assessment policies related to classroom assessment and standardized testing programs, (d) be able to draft teacher and administrator licensing requirements and training program guidelines that demand and promote competence in assessment, and (e) be able to assist districts and states to devise staffing and hiring plans assuring assessment competence in key positions in their organizational structures. Cosponsored with AERA.

QQ. Applications of Item Response Theory: Models and Methods for the Analysis of Multiple-Categorical Item Response Data
Presenters: James O. Ramsay, McGill University, and David Thissen, University of North Carolina at Chapel Hill

Sunday, April 3, 12:00 noon–6:00 p.m. and Monday, April 4, 9:00 a.m.–6:00 p.m.  Fee: $90

This session is intended to be a second course in more advanced topics than have often been included in NCME training sessions on item response theory (IRT). It is intended for persons with some basic familiarity with IRT who wish to add techniques for the analysis of multiple-categorical responses to their repertoires. By the end of the training session, the participants should be able to: (a) explain the psychological and statistical models underlying the parametric approach to multiple-categorical IRT, (b) explain the ideas underlying nonparametric approaches to multiple-categorical IRT, (c) make informed choices about which models and methods are likely to be most useful in a particular context, (d) use computer software that produces the graphical displays that comprise an IRT-based analysis of multiple-category item response data, and (e) interpret graphical displays as they relate to the content and measurement properties of the items and the response alternatives.

XX. DIF/Bias Modeling and Detection Using SIBTEST with an Emphasis on Testlet DIF/Bias
Presenter: William F. Stout, University of Illinois at Urbana-Champaign

Monday, April 4, 8:30 a.m.–4:30 p.m.  Fee: $60

The central purpose of this session is to teach both practitioners and theoreticians how to use SIBTEST effectively to detect DIF/bias. Special emphasis will be placed on the detection and interpretation of testlet DIF; that is, a clump of DIF items that combine to produce a practically important amount of DIF. Emphasis also will be placed upon demonstrating SIBTEST’s versatility: for example, it can investigate either single item or testlet DIF; it can detect DIF with good power for either dichotomously or polychronously scored items; it can detect both uniform and nonuniform DIF with good power; and it is computationally extremely fast and, hence, a thorough exploratory and/or confirmatory DIF testlet analysis is computationally feasible.

J. Performance Assessment: Issues in Development, Scoring, and Use
Presenters: Catherine J. Welch, Deborah J. Harris, Timothy R. Miller, and Edward W. Wolfe, American College Testing (ACT)

Monday, April 4, 8:00 a.m.–11:45 a.m.  Fee: $15

The primary objective of this session will be to familiarize participants with approaches to solving practical problems that are encountered in the development, scoring, and use of performance assessments. Specific topics to be addressed include: (a) development procedures that minimize topic effects, rater effects, and interactions of these effects, (b) procedures for assessing and minimizing differential item functioning in performance assessment items, (c) rater training procedures that emphasize stability over time and minimize the effects of rater variability, (d) procedures for estimating the reliability and generalizability of the scores, and (e) procedures for establishing comparability of multiple forms of the same performance assessment. Cosponsored with AERA.