APPLIED MEASUREMENT IN EDUCATION, *16*(4), 323–342
Copyright © 2003, Lawrence Erlbaum Associates, Inc.

Classroom Assessment Practices and Teachers' Self-Perceived Assessment Skills

Zhicheng Zhang
Fairfax County Public Schools

Judith A. Burry-Stock
The University of Alabama

This study investigates teachers' assessment practices across teaching levels and content areas, as well as teachers' self-perceived assessment skills as a function of teaching experience and measurement training. Data from 297 teachers on the Assessment Practices Inventory were analyzed in a MANOVA design. As grade level increases, teachers rely more on objective tests in classroom assessment and show an increased concern for assessment quality ($p < .001$). Across content areas, teachers' involvement in assessment activities reflects the nature and importance of the subjects they teach ($p < .001$). Regardless of teaching experience, teachers with measurement training report a higher level of self-perceived assessment skills in using performance measures; in standardized testing, test revision, and instructional improvement; as well as in communicating assessment results ($p < .05$) than those without measurement training. The implications of the results for measurement training are also discussed.

Classroom assessment has received increased attention from the measurement community in recent years. Since teachers are primarily responsible for evaluating instruction and student learning, there is a widespread concern about the quality of classroom assessment. Literature on classroom assessment has delineated the content domain in which teachers need to develop assessment skills (e.g., Airasian, 1994; Carey, 1994; O'Sullivan & Chalnick, 1991; Schafer, 1991; Stiggins, 1992, 1997). The current consensus has been that teachers use a variety

Requests for reprints should be sent to Zhicheng Zhang, Office of Program Evaluation, Department of Educational Accountability, Fairfax County Public Schools, Leis Instructional Center, 7423 Camp Alger Ave., Falls Church, VA 22042. E-mail: zzhang@fcps.edu

of assessment techniques, even though they may be inadequately trained in certain areas of classroom assessment (Hills, 1991; Nolen, Haladyna, & Haas, 1992; Plake, 1993; Stiggins & Conklin, 1992). Less researched, however, is how teachers perceive their assessment practices and assessment skills. This study seeks to expand the current research on classroom assessment by examining teachers' assessment practices and self-perceived assessment skills in relation to content area, grade level, teaching experience, and measurement training.

RELATED LITERATURE

Classroom Assessment

Classroom assessment embraces a broad spectrum of activities from constructing paper-pencil tests and performance measures, to grading, interpreting standardized test scores, communicating test results, and using assessment results in decision-making. When using paper-pencil tests and performance measures, teachers should be aware of the strengths and weaknesses of various assessment methods, and choose appropriate formats to assess different achievement targets (Stiggins, 1992). Test items should match with course objectives and instruction to ensure content validity (Airasian, 1994), reflect adequate sampling of instructional materials to improve test reliability, and tap higher-order thinking skills. In performance assessment, validity and reliability can be improved by using observable and clearly defined performance tasks (Airasian, 1994; Baron, 1991; Shavelson, Baxter, & Pine, 1991; Stiggins, 1987), detailed scoring protocols, multiple samples of behaviors evaluated by several judges (Dunbar, Koretz, & Hoover, 1991), and recording scoring results during assessment (Stiggins & Bridgeford, 1985). Teachers should be able to revise and improve teacher-made tests based on test statistics and item analysis (Carey, 1994; Gregory, 1996).

Grading and standardized testing are two important components of classroom assessment. Since grade-based decisions may have lasting academic and social consequences (Messick, 1989; Popham, 1997), teachers should weigh assessment components according to instructional emphasis (Airasian, 1994; Carey, 1994; Stiggins, Frisbie, & Griswold, 1989) and base grades on achievement-related factors only. Grading criteria should be communicated to students in advance and implemented systematically to handle regular as well as borderline cases (Stiggins et al., 1989). Nonachievement factors such as effort, ability, attitude, and motivation should not be incorporated into subject-matter grades because they are hard to define and measure (Stiggins et al., 1989). In terms of standardized testing, teachers should avoid teaching to the test (Mehrens, 1989), interpreting test items, and giving hints or extra time during test administration. Teachers should appropriately interpret test scores and identify diagnostic information from test results about instruction and student learning (Airasian, 1994).

Communicating assessment results and using assessment information in decision-making constitute two other aspects of classroom assessment. To communicate assessment results effectively, teachers must understand the strengths and limitations of various assessment methods, and be able to use appropriate assessment terminology and communication techniques (Schafer, 1991; Stiggins, 1997). Specific comments rather than judgmental feedback (e.g., "fair") are recommended to motivate students to improve performance (Brookhart, 1997). When using assessment results, teachers should protect students' confidentiality (Airasian, 1994). Teachers should also be able to use assessment results to make decisions about students' educational placement, promotion, and graduation, as well as to make judgment about class and school improvement (Stiggins, 1992).

In 1990, the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) issued Standards for Teacher Competence in Educational Assessment of Students. These standards are currently under revision. According to the standards, teachers should be skilled in choosing and developing assessment methods, administering and scoring tests, interpreting and communicating assessment results, grading, and meeting ethical standards in assessment. The assessment literature and the seven standards form the theoretical framework for the investigation of teachers' assessment practices and skills in this study.

Teachers' Assessment Practices and Competencies

Investigations of teachers' assessment practices revealed that teachers were not well prepared to meet the demand of classroom assessment due to inadequate training (Goslin, 1967; Hills, 1991; O'Sullivan & Chalnick, 1991; Roeder, 1972). Problems were particularly prominent in performance assessment, interpretation of standardized test results, and grading procedures. When using performance measures, many teachers did not define levels of performance or plan scoring procedures before instruction, nor did they record scoring results during assessment (Stiggins & Conklin, 1992). In terms of standardized testing, teachers reported having engaged in teaching test items, increasing test time, giving hints, and changing students' answers (Hall & Kleine, 1992; Nolen, Haladyna, & Haas, 1992). Teachers also had trouble interpreting standardized test scores (Hills, 1991; Impara, Divine, Bruce, Liverman, & Gay, 1991) and communicating test results (Plake, 1993). Many teachers incorporated nonachievement factors such as effort, attitude, and motivation into grades (Griswold, 1993; Hills, 1991; Jongsma, 1991; Stiggins et al., 1989) and they often did not apply weights in grading to reflect the differential importance of various assessment components (Stiggins et al., 1989). Despite the aforementioned problems, most teachers believed that they had adequate knowledge of testing (Gullikson, 1984; Kennedy, 1993) and attributed that knowledge to experience and university coursework (Gullikson, 1984; Wise, Lukin, & Roos, 1991).

Teachers' concern about the quality of classroom assessment varied with grade levels and slightly with subject areas (Stiggins & Conklin, 1992). There was an increased concern among teachers about the improvement of teacher-made objective tests at higher-grade levels; mathematics and science teachers were more concerned about the quality of the tests they produced than were writing teachers. Higher-grade level mathematics teachers were found to attach more importance to and use more frequently homework and teacher-made tests in classroom assessment than lower-grade level teachers (Adams & Hsu, 1998).

Two points are noteworthy about the existing literature. First, assessment practices and assessment skills are related but have different constructs. Whereas the former pertains to assessment activities, the latter reflects an individual's perception of his or her skill level in conducting those activities. This may explain why teachers rated their assessment skills as good even though they were found inadequately prepared to conduct classroom assessment in several areas. Current literature is scarce in simultaneous investigation of assessment practices and assessment-related perceptions. Second, classroom assessment involves a broad range of activities. Teachers may be involved in some activities more than in others due to the nature of assessment specific to the grade levels and content areas they are required to teach. Although the existing literature has suggested that grade levels and subject areas may account for some variations in classroom assessment (Adams & Hsu, 1998; Stiggins & Conklin, 1992), none of these studies, however, have covered sufficiently the broad spectrum of classroom assessment. Further research addressing teachers' assessment practices and their self-perceived assessment skills in various assessment activities in light of teaching levels and content areas is desirable to strengthen the current literature on classroom assessment. These two points provide the rationale for this study.

The primary purpose of this study was to investigate teachers' assessment practices and self-perceived assessment skills. Specifically, this study aimed at achieving three objectives: (1) to investigate the relationship between teachers' assessment practices and self-perceived assessment skills, (2) to examine classroom assessment practices across teaching levels and content areas, and (3) to examine teachers' self-perceived assessment skills in relation to years of teaching and measurement training. Embedded in these objectives is the premise that assessment practices are impacted by content and intensity of instruction whereas self-perceived assessment skills are influenced mainly by teaching experience and professional training (Gullikson, 1984).

METHOD

Instrument

An Assessment Practices Inventory (API) (Zhang & Burry-Stock, 1994) was used in this study. The instrument was developed within the theoretical framework delineated by the literature on classroom assessment (e.g., Airasian, 1994; Carey,

1994; O'Sullivan & Chalnick, 1991; Schafer, 1991; Stiggins, 1991) and the Standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, & NEA, 1990). To ensure the content validity of the instrument, a table of specifications was used to generate items for each major aspect of classroom assessment. All together 67 items were produced to cover a broad range of assessment activities including constructing paper-pencil tests and performance measures, interpreting standardized test scores, grading, communicating assessment results, and using assessment results in decision-making. The instrument was piloted twice with inservice teachers and revisions were made based on the teachers' feedback and item analyses (Zhang, 1995).

Teachers were asked to mark their responses to the same 67 items on two different rating scales: use scale and skill scale. The use scale was designed to measure teachers' *assessment practices* with the following scale ranging from 1 (*not at all used*), 2 (*seldom used*), 3 (*used occasionally*), 4 (*used often*) to 5 (*used very often*). The skill scale was designed to measure teachers' *self-perceived assessment* skills with its scale points ranging from 1 (*not at all skilled*), 2 (*a little skilled*), 3 (*somewhat skilled*), 4 (*skilled*) to 5 (*very skilled*). Thus, two data sets were produced, with one on assessment practices and the other on self-perceived assessment skills. The items of the API are presented in Appendix A.

Sample and Procedure

The API was sent to the entire instructional work force of 845 teachers in two school districts in a southeastern state. One school district was predominantly rural and suburban and the other predominantly urban. The numbers of elementary, middle, and high schools participating in this study were 6, 4, and 6, respectively. The instrument, together with a cover letter, and a computer scanable answer sheet were distributed to the teachers by their principal at faculty meetings. Those who voluntarily responded to the survey returned the completed answer sheets to the school secretary. The answer sheets were collected by the first author.

Two hundred and ninety-seven completed surveys were used in the analyses. The teachers responding to the survey were predominantly white (89%) and female (77.4%). Two hundred and sixty-three respondents clearly classified themselves into one of the three teaching levels: elementary school, 38.8%; middle school, 28.5%; and high school, 32.7%. One hundred and fifty-four respondents reported teaching one content area whereas others taught multiple subjects such as language arts, math, and physical education. The distribution of single-subject respondents was as follows: language arts, 25.3%; mathematics, 24%; science, 16.2%; social studies, 14.9%; and nonacademic subjects (arts, home economics, keyboard, music, and physical education), 19.5%. Most respondents (96%) had a bachelor's degree; 56% held a Master's degree. About 82% of the teachers had had at least one measurement course. Table 1 presents summary information on respondents by teaching level and content area.

TABLE 1
Teacher Information by Teaching Level and Content Area

Teaching Level	*n*[a]	*Content Area*	*n*[a]
Elementary school	102	Language arts	39
Middle school	75	Mathematics	37
High school	86	Science	25
Others[b]	28	Social studies	23
		Nonacademic subjects[c]	30
		Multiple subjects[d]	122

Note. [a]n may not add up to 297 due to missing data. [b]Refer to those teaching comprehensive k-8, comprehensive k-12, and undefined levels. [c]Refer to arts, keyboard, home economics, music, and physical education. [d]Apply to those teaching multiple academic subjects.

RESULTS

Factor Structures of Assessment Practices and Self-Perceived Assessment Skills

The data sets on assessment practices and self-perceived assessment skills were factor-analyzed with a principal axis method of extraction and a varimax orthogonal rotation. Principal factor analysis was used because it provides more information and is easier to interpret than principal component analysis (Cureton & D'Agostino, 1983). Given that assessment practices are mostly behavior-based whereas self-perceived assessment skills reflect behavior-based perception, it was hypothesized that the underlying dimensions of the two constructs would overlap to some extent. The 67 items converged on six factors for assessment practices and seven factors for self-perceived assessment skills based on the screen plot and eigenvalues for the initial solution (the first six eigenvalues for assessment practices were: 13.15, 5.24, 3.56, 2.70, 2.25, and 1.73; the first seven eigenvalues for self-perceived assessment skills were: 21.06, 4.20, 2.67, 2.40, 2.15, 1.63, and 1.24). The six factors for assessment practices were: (1) Using Paper–Pencil Tests; (2) Standardized Testing, Test Revision, and Instructional Improvement; (3) Communicating Assessment Results, Ethics, Grading; (4) Using Performance Assessment; (5) Nonachievement-Based Grading; and (6) Ensuring Test Reliability and Validity. The seven factors for self-perceived assessment skills were: (1) Perceived Skillfulness in Using Paper-Pencil Tests; (2) Perceived Skillfulness in Standardized Testing, Test Revision, and Instructional Improvement; (3) Perceived Skillfulness in Using Performance Assessment; (4) Perceived Skillfulness in Communicating Assessment Results; (5) Perceived Skillfulness in Nonachievement-Based Grading; (6) Perceived Skillfulness in Grading and Test Validity; and (7) Perceived Skillfulness in Addressing Ethical Concerns. The factor that

pertains to nonachievement-based grading in both cases (i.e., factor 5 for assessment practices and factor 5 for self-perceived assessment skills) subsumes items describing the practice of grading on attendance, ability, effort, and behavior. Given that grading on nonachievement-related factors is not recommended by measurement experts, low scores on the items related to this factor are preferred than high scores. The amount of the variance explained by the factor structure was 42.8% and 52.8% for assessment practices and self-perceived assessment skills, respectively. The percent of the variance explained by the individual factors ranged from 11.2 to 4.1 for assessment practices and from 12.2 to 3.9 for self-perceived assessment skills. Table 2 presents the information on the number of items, the percent of the variance explained, the range of item loadings, and Cronbach alpha reliability for each factor for both assessment practices and self-perceived assessment skills. Even though the total scores of assessment practices and self-perceived assessment skills were not used in this study, the Cronbach alpha reliability coefficients are reported for the overall scales in Table 2 for future reference.

TABLE 2
Factor Structures of Assessment Practices and Self-Perceived Assessment Skills (N = 297)

Assessment Practices				*Self-Perceived Assessment Skills*			
Factors	*Number of Items*	*Variance*	*Reliability*	*Factors*	*Number of Items*	*Variance*	*Reliability*
UPP	19	11.2 (.74 – .29)	.89	PSPP	16	12.2 (.75 – .36)	.91
STRI	14	8.6 (.72 – .35)	.87	PSSTRI	14	10.2 (.72 – .46)	.91
COMEG	15	8.5 (.56 – .36)	.87	PSPA	10	8.8 (.77 – .42)	.89
UPA	9	6.1 (.75 – .33)	.82	PSCOM	9	6.6 (.65 – .36)	.87
NG	5	4.3 (.76 – .47)	.80	PSNG	6	5.8 (.74 – .30)	.85
ETVR	5	4.1 (.58 – .44)	.77	PSGTV	10	5.3 (.50 – .38)	.88
				PSET	2	3.9 (.73 – .72)	.90
Total	67	42.8	.94		67	52.8	.97

Note. For Assessment Practices: UPP = Using Paper–Pencil Tests; STRI = Standardized testing, Test Revision, and Instructional Improvement; COMEG = Communicating Assessment Results, Ethics, and Grading; UPA = Using Performance Assessment; NG = Nonachievement-Based Grading; ETVR = Ensuring Test Validity and Reliability. For Self-Perceived Assessment Skills: PSPP = Perceived Skillfulness in Using Paper–Pencil Tests; PSSTRI = Perceived Skillfulness in Standardized testing, Test Revision, and Instructional Improvement; PSPA = Perceived Skillfulness in Using Performance Assessment; PSCOM = Perceived Skillfulness in Communicating Assessment Results; PSNG = Perceived Skillfulness in Nonachievement-Based Grading; PSGTV = Perceived Skillfulness in Grading and Test Validity; PSET = Perceived Skillfulness in Addressing Ethical Concerns. Variance = percent of the variance explained after rotation. The numbers in parenthesis indicate the range of item loadings. Given these reliability coefficients were derived from the same data that were factor analyzed to form the scales, please keep in mind that the reliabilities reported here may be inflated.

The similarities and differences between the factor structures of assessment practices and self-perceived assessment skills may be highlighted with two points. First, with the exception of one or two items, the two factor structures were very similar in the four underlying dimensions that address paper-pencil tests; standardized testing, test revision, and instructional improvement; performance assessment; and nonachievement-based grading. In other words, factors 1, 2, 4, and 5 on assessment practices correspond fairly well with factors 1, 2, 3,and 5 on self-perceived assessment skills, respectively. Second, although Perceived Skillfulness in Communicating Assessment Results, Perceived Skillfulness in Addressing Ethical Concerns, and Perceived Skillfulness in Grading and Test Validity emerged as three distinct factors for self-perceived assessment skills, most of the same items converged on two factors for assessment practices with one embodying the items related to communication, grading, and ethics (factor 3 for assessment practices) and the other subsuming the items on test validity and reliability (factor 6 for assessment practices). Since communicating assessment results always involves reporting grades and the ethical issue of protecting students' confidentiality, this finding seems to suggest that the construct of assessment practices captures the internal connections among different assessment activities more than that of self-perceived assessment skills. Lending support to this suggestion is the finding that the items that converged on the factor of Ensuring Test Reliability and Validity (factor 6 for assessment practices) were initially written for different assessment activities pertaining to assessment planning (i.e., developing assessments based on clearly defined course objectives), developing paper-pencil tests (i.e., ensuring adequate content sampling for a test), and using performance assessment (i.e., defining a rating scale for performance criteria in advance, matching performance tasks to instruction and objectives). Once again, assessment practices as a construct captured what these items have in common and subsumed the items under the same dimension. This finding provides additional evidence for our conclusion that assessment practices as a construct is more coherent than self-perceived assessment skills. Appendix B indicates where the items load on the two factor structures.

Overall, these findings confirm our hypothesis that the constructs of assessment practices and self-perceived assessment skills overlap to some extent in terms of the underlying dimensions they measure, yet each construct maintains a certain degree of uniqueness. The overlap between assessment practices and self-perceived assessment skills was also reflected in a Pearson product–moment correlation coefficient of .71 that explained 50% of the shared variance between the two constructs. The difference between assessment practices and self-perceived assessment skills mainly stems from the fact that the former is largely behavior-based and thus internally coherent, whereas the latter reflects teachers' perception of their ability to perform classroom assessment and, as a result, less predictable. Based on factor analyses, six and seven scales were formed for assessment

practices and self-perceived assessment skills, respectively. Each scale score was generated by summing up the individual scores of the items loading high on a factor. As reported in Table 2, the Cronbach alpha reliabilities of the scales ranged from .89 to .77 for assessment practices and from .91 to .85 for self-perceived assessment skills. Given that these reliabilities were derived from the same data that were factor-analyzed to form the scales, it should be noted that these reliabilities may be inflated.

Assessment Practices Across Teaching Levels and Content Areas

The scale scores for assessment practices were used in two separate one-way multivariate analyses: one by teaching levels (elementary, middle, and high school) and the other by content areas (language arts, mathematics, science, social studies, and nonacademic subjects that include arts, home economics, key board, music, and physical education). Significant multivariate main effects were revealed for teaching levels, $F(12, 510) = 7.95$, $p < .001$ (Wilks' Lambda = .71). As reported in Table 3, significant differences were observed across teaching levels in Using Paper-Pencil tests, $F(2, 260) = 17.80$, $p < .001$; Using Performance Assessment, $F(2, 260) = 4.90$, $p < .01$; and Ensuring Test Reliability and Validity,

TABLE 3
Two One-Way MANOVAs: Assessment Practices by Teaching Level and Content Area

	Teaching Level (N = 263)				*Content Area (N = 154)*					
Scale	*Elem*	*Middle*	*High*	*F*	*LA*	*MA*	*SC*	*SS*	*NA*	*F*
UPP	59.56	69.92	70.28	17.80***	73.49	64.76	71.28	71.74	66.50	4.00**
STRI	37.99	35.33	35.90	1.46	39.13	36.68	31.92	36.65	30.50	3.59**
COMEG	57.80	57.60	59.33	.77	60.05	61.51	57.36	56.22	54.50	2.94*
UPA	34.28	31.28	31.79	4.90**	32.08	30.11	29.16	30.61	33.47	1.61
NG	15.66	15.40	15.34	.11	15.26	17.14	17.64	14.48	14.47	2.72*
ETVR	16.99	19.32	19.77	12.01***	19.95	19.35	17.92	19.13	19.43	1.06

Note. UPP = Using Paper–Pencil Tests; STRI = Standardized testing, Test Revision, and Instructional Improvement; COMEG = Communicating Assessment Results, Ethics, and Grading; UPA = Using Performance Assessment; NG = Nonachievement-Based Grading; ETVR = Ensuring Test Validity and Reliability. Elem. = elementary teachers; Middle = middle school teachers; High = high school teachers. LA = language arts; MA = mathematics; SC = science; SS = social studies; NA = arts/keyboard/home economics/music/physical education. For teaching levels: Wilks' Lambda = .71 and multivariate F value = 7.95, $p < .001$. For content areas: Wilks' Lambda = .58 and multivariate F value = 3.59, $p < .001$. Unless specified otherwise, the numbers represent mean scale scores.

*$p < .05$. **$p < .01$. ***$p < .001$.

$F(2, 260) = 12.01$, $p < .001$. Middle school and high school teachers used paper–pencil tests more often than did elementary school teachers; the former also used the recommended techniques to ensure test reliability and validity more often than the latter. Elementary school teachers used performance assessment more often than did middle school and high school teachers. These results suggest a general distinction between elementary and secondary teachers. Whereas secondary teachers rely on teacher-made objective tests more often and are more concerned about the quality of classroom assessment, elementary teachers use alternative measures more often to assess student learning.

Only teachers who reported teaching one subject area were used in the MANOVA analysis for content areas. As can be seen from Table 3, significant multivariate main effects were revealed for content areas, $F(24, 503) = 3.59$, $p < .001$ (Wilks' Lambda = .58). Significant differences were noticed across content areas in Using Paper-Pencil Tests, $F(4, 149) = 4.00$, $p < .01$; Standardized Testing, Test Revision, and Instructional Improvement, $F(4, 149) = 3.59$, $p < .01$; Communicating Assessment Results, Ethics, and Grading, $F(4, 149) = 2.94$, $p < .05$; and Nonachievement-Based Grading, $F(4, 149) = 2.72$, $p < .05$. Teachers in language arts, science, and social studies used paper–pencil tests more often than did mathematics teachers; language-arts teachers used paper-pencil tests more often than did teachers in nonacademic subjects. Teachers in language arts, mathematics, and social studies were engaged in interpreting standardized tests, revising tests, and improving instruction based on assessment results more often than those teaching nonacademic subjects; language-arts teachers were also engaged in interpreting standardized tests, revising tests, and improving instruction based on assessment results more often than science teachers. Mathematics and language-arts teachers reported more frequent use of assessment activities of communicating assessment results, meeting ethical standards, and grading than did nonacademic-subjects teachers; mathematics teachers also reported more frequent involvement in these assessment activities than did social-studies teachers. Finally, mathematics and science teachers reported grading on nonachievement-related factors more frequently than did teachers in social studies and nonacademic subjects, possibly suggesting teachers' belief that motivation and efforts have an impact on achievement in demanding subjects such as mathematics and science despite recommendation against such grading practices in the measurement community. These findings seem to suggest that content-related variations in teachers' assessment practices are largely reflective of the nature and relative importance of the various subjects taught at school. Overall, teachers in academic subjects are involved in certain assessment activities (e.g., interpreting standardized test results and communicating assessment results) more often than those teaching nonacademic subjects because academic subjects such as mathematics, language arts, science, and social studies are often covered on statewide standardized tests (e.g., VDE, 1998).

Assessment Skills by Years of Teaching and Measurement Training

Do teachers' self-perceived assessment skills vary with teaching experience? How does measurement training contribute to teachers' self-perceived assessment skills? To answer these questions, a 3-by-2 MANOVA was performed on the assessment-skills data to examine teachers' self-perceived assessment skills as a function of years of teaching and measurement training. The seven scale scores generated out of the factor analysis were used as the dependent variables for self-perceived assessment skills. There are two independent variables: years of teaching (≦1 year, 2–5 years, and ≧ 6 years) and measurement training (no measurement training, at least one measurement course).

In Table 4 the following statistics are summarized. Each of the scales for self-perceived assessment skills is listed in column 1 with the number of items associated with each scale being presented in column 2. The mean scale scores are

TABLE 4
A 3-by-2 MANOVA: Self-Perceived Assessment Skills as a Function of Years of Teaching and Measurement Training (N = 297)

Scale	*Number of Items*	*Measurement Training*	*Years of Teaching*			*F*
			≤ 1 Year	*2–5 Years*	*≥ 6 Years*	
PSPP	16	No Training	47.33	58.25	59.15	
		At Least One Course	58.89	64.37	62.05	
PSSTRI	14	No Training	29.00	38.75	35.70	17.54***
		At Least One Course	43.29	42.40	42.88	
PSPA	10	No Training	30.33	36.25	34.68	9.01**
		At Least One Course	37.32	40.53	37.76	
PSCOM	9	No Training	26.00	34.75	33.20	6.45*
		At Least One Course	33.24	36.73	35.34	
PSNG	6	No Training	17.00	20.50	20.70	
		At Least One Course	18.53	21.53	20.17	
PSGTV	10	No Training	28.67	35.25	37.15	
		At Least One Course	36.95	39.63	38.13	
PSET	2	No Training	6.67	7.75	7.80	
		At Least One Course	7.74	7.67	8.04	

Note. PSPP = Perceived Skillfulness in Using Paper–Pencil Tests; PSSTRI = Perceived Skillfulness in Standardized testing, Test Revision, and Instructional Improvement; PSPA = Perceived Skillfulness in Using Performance Assessment; PSCOM = Perceived Skillfulness in Communicating Assessment Results; PSNG = Perceived Skillfulness in Nonachievement-Based Grading; PSGTV = Perceived Skillfulness in Grading and Test Validity; PSET = Perceived Skillfulness in Addressing Ethical Concerns. For multivariate main effects for measurement training: Wilks' Lambda = .94 and F value = 2.50, $p < .05$. $F = F$ value for univariate main effects for measurement training. Only significant F values are reported. Unless specified otherwise, the numbers represent mean scale scores.

*$p < .05$. **$p < .01$. ***$p < .001$.

presented for people with different years of teaching experience (columns 4–6) and measurement training background (two adjacent rows for each scale in column 3). For example, the mean scale score on Perceived Skillfulness in Using Paper–Pencil Tests is 47.33 for teachers with no measurement training and with ≤1 year of teaching experience whereas the mean scale score for those with the same limited teaching experience but with measurement training (at least one course) is 58.89. The *F* values reported in column 7 indicate significant differences between means (see further discussion given later).

As can be seen in Table 4, significant multivariate main effects were found for measurement training $F(7, 275) = 2.50$, $p < .05$ (Wilks' Lambda = .94). Further examination of univariate main effects revealed that teachers who had received measurement training perceived themselves to be more skilled than those without measurement training in standardized testing, test revision, and instructional improvement, $F(6, 281) = 17.54$, $p < .001$; in using performance assessment, $F(6, 281) = 9.01$, $p < .01$; and in communicating assessment results, $F(6, 281) = 6.45$, $p < .05$. No significant main effects were detected for years of teaching, nor were there significant interaction effects between measurement training and years of teaching. These results seem to suggest that, regardless of teaching experience, teachers' self-perceived assessment skills are augmented by measurement training. Overall, these findings testify the value of measurement training.

SUMMARY AND CONCLUSIONS

This study investigates teachers' assessment practices and self-perceived assessment skills within the framework of classroom assessment literature and the Standards for Teacher Competence in Educational Assessment of Students. Factor analytical technique was applied to study the relationship between the constructs of assessment practices and self-perceived assessment skills. Teachers' assessment practices and self-perceived assessment skills were examined in a MANOVA design to determine how they may vary as a function of teaching level, content area, teaching experience, and measurement training.

The constructs of assessment practices and self-perceived assessment skills overlap to some extent in the underlying dimensions they measure, yet each contains a certain degree of uniqueness. The similarity between assessment practices and self-perceived assessment skills is supported by a strong correlation coefficient of .71 and by similar patterns of item loadings on four of the underlying dimensions they measure (paper-pencil test; standardized testing, test revision, and instructional improvement; performance assessment; and nonachievement-based grading). Where the two factor structures differ, the construct of assessment practices does a better job of subsuming inherently related activities under the same dimension than does that of self-perceived assessment skills. The finding that the

factors on assessment practices are more reflective of the intrinsic nature of classroom assessment activities than those on self-perceived assessment skills adds credence to our view that the former is more inherently coherent than the latter.

Teachers differ in their assessment practices due to the nature of classroom assessment delineated by teaching levels. A general difference emerges between elementary and secondary teachers in terms of the assessment methods used and teachers' concerns for assessment quality. While secondary teachers rely mostly on paper–pencil tests and were concerned about the quality of assessment, elementary teachers often use performance assessment as an alternative. These results confirm the previous research findings that, as grade level increases, teachers rely more on objective techniques in classroom assessment and show an increased concern for assessment quality (Adams & Hsu, 1998; Stiggins & Conklin, 1992). Whereas frequent use of objective tests at the secondary level may have occurred as a result of teachers' needs to tailor tests to cover unique classroom objectives at higher-grade levels, the increased concern about assessment quality at secondary level is reflective of the fact that grades and assessment-based decisions take on more importance as students progress in the school system (Stiggins & Conklin, 1992).

The results of this study also lend support to the conclusion that teachers' assessment practices differ across content areas (Gullikson, 1984; Stiggins & Conklin, 1992; Zhang, 1995). The variations emerging from this study indicate that teachers' assessment practices are driven by the subjects they teach. The finding also implies a greater need to interweave measurement training with content areas. Inservice teachers enrolled in a measurement course should be encouraged to base their assessment projects on the instructional activities taking place in their own classrooms. For preservice teachers, assessment projects should be integrated with student teaching and other practical experiences.

Knowledge in measurement and testing has a significant impact on teachers' self-perceived assessment skills regardless of their teaching experience. This is particularly true in terms of interpreting standardized test scores, revising teacher-made tests, modifying instruction based on assessment feedback; using performance assessment; and communicating assessment results. This finding confirms teachers' beliefs that university coursework contributes to their knowledge of testing and measurement (Gullikson, 1984; Wise, Lukin, & Roos, 1991). It also implies that measurement training may compensate for novices' lack of experience in the classroom. Previous research has indicated that teachers had trouble interpreting standardized test results (Impara et al., 1991), that they were inadequate in defining and scoring performance measures (Stiggins & Conkin, 1992), and that they were not proficient in communicating assessment results (Plake, 1993). Yet, it is in these areas of classroom assessment that measurement training enhances teachers' self-perceived assessment skills. The results of this study provide evidence for the value of university coursework in tests and measurement.

The generalizability of the specific results of this study may be limited by its use of self-report surveys and the participating sample. Future studies may use multiple methods of data collection including classroom observation, analysis of teacher-made tests, teachers' grade books, and teacher interviews to validate teacher self-reports. In the future, the survey should be sent to a more representative sample selected from a variety of geographic regions across the country.

The results of this study have implications for both researchers and teacher educators. From a research perspective, the association between teachers' assessment practices and self-perceived assessment skills revealed in this study is intriguing and calls for further exploration. Future research may focus on the relationship between teachers' assessment practices and assessment skills, particularly on how assessment skills may facilitate improvement in assessment practices and on what practices are prerequisite to developing assessment skills.

This study supports the value of measurement training. The finding that teachers' assessment practices are reflective of the nature of the subjects and grade levels they teach implies that measurement training programs should be tailored to suit the deferential needs of teachers working in different content areas and grade levels. One way to achieve this is to encourage inservice teachers to base their measurement projects on the instructional and assessment activities taking place in their own classrooms (Taylor, 1997). This approach has been implemented with encouraging results in a staff development program designed to facilitate changes in mathematics teachers' assessment practices (Borko, Mayfield, Marion, Flexer, & Cumbo, 1997). For preservice teachers who have limited classroom experience to draw on during measurement training, modeling assessment practices through course projects may be an effective instructional approach in teacher preparation coursework (Burry-Stock & Cochran, 1995; Criswell & Criswell, 1995). These hands-on projects and real-life examples may facilitate teachers to transfer measurement knowledge into assessment practices in their own classrooms (Phye, 1992). Assessment is the feedback mechanism for improving classroom learning. By improving teachers' assessment practices and skills, we can improve classroom learning. This is an ambitious task, but herein lies a way of improving student achievement.

ACKNOWLEDGMENTS

The authors thank Dr. Weiyun Chen and two anonymous reviewers for their valuable feedback and suggestions for revising this article.

REFERENCES

Adams, E. L., & Hsu, J. Y. (1998). Classroom assessment: Teachers' conceptions and practices in mathematics. *School Science and Mathematics, 98*(4), 174–180.

Airasian, P. W. (1994). *Classroom assessment*. New York: McGraw-Hill.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association (AFT, NCME, & NEA). (1990). *Standards for teacher competence in educational assessment of students*. Washington, DC: National Council on Measurement in Education.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 305–318.

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Success, stumbling blocks, and implications for professional development. *Teaching & Teacher Education, 13*(3), 259–278.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education, 10*(2), 161–180.

Burry-Stock, J. A., & Cochran, H. K. (1995). *Assessment of Classroom Learning: BER450 Handbook*. Tuscaloosa, AL: The University of Alabama.

Carey, L. M. (1994). *Measuring and evaluating school learning*. Boston: Allyn and Bacon.

Criswell, J. R., & Criswell, S. J. (1995). Modeling alternative classroom assessment practices in teacher education coursework. *Journal of Instructional Psychology, 22,* 190–193.

Cureton, E. E., & D'Agostino, R. B. (1983). *Factor analysis: An applied approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303.

Goslin, D. A. (1967). *Teachers and testing*. New York: Russell Sage Foundation.

Gregory, R. J. (1996). *Psychological testing: History, principles, and applications* (2nd ed.). Boston: Allyn and Bacon.

Griswold, P. A. (1993). Beliefs and inferences about grading elicited from student performance sketches. *Educational Assessment, 1*(4), 311–328.

Gullikson, A. R. (1984). Teacher perspectives of their instructional use of tests. *Journal of Educational Research, 77*(4), 244–248.

Hall, J. L., & Kleine, P. F. (1992). Educators' perceptions of NRT misuse. *Educational Measurement: Issues and Practices, 11*(2), 18–22.

Hills, J. R. (1991). Apathy concerning grading and testing. *Phi Delta Kappa, 72*(7), 540–545.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice, 10*(4), 16–18.

Jongsma, K. S. (1991). Rethinking grading practices. *The Reading Teacher, 45*(4), 319–320.

Kennedy, E. (1993). Evaluation of classroom assessment practices: Practitioner criteria. *College Student Journal, 27,* 342–345.

Mehrens, W. A. (1989). *Preparing students to take standardized achievement tests*. (Digest EDO-TM-89-2). Washington, DC: American Institution for Research.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice, 11*(2), 9–15.

O'Sullivan, R. G., & Chalnick, M. K. (1991). Measurement-Related course work requirements for teacher certification and recertification. *Educational Measurement: Issues and Practices, 10*(1), 17–19.

Phye, G. D. (1992). Strategic transfer: A tool for an academic problem solving. *Educational Psychology Review, 4,* 393–421.

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6*(1), 21–27.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.

Roeder, H. H. (1972). Are today's teachers prepared to use tests? *Peabody Journal of Education, 59,* 239–240.

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice, 10*(1), 3–6.
Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347–362.
Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practices, 16*(3), 33–42.
Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practices, 10*(1), 7–12.
Stiggins, R. J. (1992). High quality classroom assessment: What does it really mean? *Educational Measurement: Issues and Practice, 11*(2), 35–39.
Stiggins, R. J. (1997). *Student-centered classroom assessment* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall
Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271–286.
Stiggins, R. J, & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany: State University of New York Press.
Stiggins, R. J., Frisbie, R. J., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice 8*(2), 5–14.
Taylor, C. S. (1997). Using portfolios to teach teachers about assessment: How to survive. *Educational Assessment, 4*(2), 123–147.
Virginia Department of Education (VDE). (1998). *Virginia standards of learning assessments: End of course.* Richmond, VA: The Commonwealth of Virginia Department of Education.
Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*(1), 37–42.
Zhang, Z. (1995). *Investigating teachers' self-perceived assessment practices and assessment competencies on the Assessment Practices Inventory*. Unpublished doctoral dissertation. Tuscaloosa, AL: The University of Alabama.
Zhang, Z., & Burry-Stock, J. A. (1994). *Assessment Practices Inventory*. Tuscaloosa, AL: The University of Alabama.

APPENDIX A

Assessment Practices Inventory (8.0)

Directions: This inventory contains 67 items that address issues in classroom assessment of student learning. For each item, please use the following rating scales to indicate (1) how frequently you use the assessment practice described by the item and (2) how skilled you are in using that assessment practice. The two rating scales are defined as follows:

Use Scale: 1 = not at all used, 2 = seldom used, 3 = used occasionally, 4 = used often, and 5 = used very often

Skill Scale: 1 = not at all skilled, 2 = a little skilled, 3 = somewhat skilled, 4 = skilled, and 5 = very skilled

1. Choosing appropriate assessment methods for instructional decisions.
2. Selecting textbook-provided test items for classroom assessment.
3. Revising previously produced teacher-made tests to match current instructional emphasis.

4. Administering announced quizzes.
5. Administering unannounced quizzes.
6. Evaluating oral questions from students.
7. Assessing students through observation.
8. Determining if a standardized achievement test is valid for classroom assessment.
9. Using a table of specifications to plan assessments.
10. Developing assessments based on clearly defined course objectives.
11. Matching assessments with instruction.
12. Writing paper-pencil tests.
13. Writing multiple-choice questions.
14. Writing matching questions.
15. Writing true/false questions.
16. Writing fill-in-the-blank or short answer questions.
17. Writing essay questions.
18. Writing test items for higher cognitive levels.
19. Constructing a model answer for scoring essay questions.
20. Ensuring adequate content sampling for a test.
21. Matching performance tasks to instruction and course objectives.
22. Defining a rating scale for performance criteria in advance.
23. Communicating performance assessment criteria to students in advance.
24. Recording assessment result on the rating scale/checklist while observing a student's performance.
25. Using concept mapping to assess student learning.
26. Assessing individual class participation.
27. Assessing group class participation.
28. Assessing individual hands-on activities.
29. Assessing group hands-on activities.
30. Assessing individual class participation.
31. Using portfolios to assess student progress.
32. Following required procedures (time limit, no hints, no interpretation) when administering standardized tests.
33. Interpreting standardized test scores (e.g., Stanine, Percentile Rank) to students and parents.
34. Interpreting Percentile Band to students and parents.
35. Calculating and interpreting central tendency and variability for teacher-made tests.
36. Conducting item analysis (i.e., difficulty and discrimination indices) for teacher-made tests.
37. Revising a test based on item analysis.
38. Obtaining diagnostic information from standardized tests.
39. Using assessment results when planning teaching.
40. Using assessment results when developing curriculum.

41. Using assessment results when making decisions (e.g., placement, promotion) about individual students.
42. Using assessment results when evaluating class improvement.
43. Using assessment results when evaluating school improvement.
44. Developing systematic grading procedures.
45. Developing a grading philosophy.
46. Using norm-referenced grading model.
47. Using criteria-referenced grading model.
48. Using systematic procedures to determine borderline grades.
49. Informing students in advance how grades are to be assigned.
50. Establishing student expectations for determining grades for special education students.
51. Weighing differently projects, exams, homework, etc. when assigning semester grades.
52. Incorporating extra credit activities in the calculation of grades.
53. Incorporating ability in the calculation of grades.
54. Incorporating classroom behavior in the calculation of grades.
55. Incorporating improvement in the calculation of grades.
56. Incorporating effort in the calculation of grades.
57. Incorporating attendance in the calculation of grades.
58. Assigning grades.
59. Providing oral feedback to students.
60. Providing written feedback to students.
61. Communicating classroom assessment results to students.
62. Communicating classroom assessment results to parents.
63. Communicating classroom assessment results to other educators.
64. Avoiding teaching to the test when preparing students for tests.
65. Protecting students' confidentiality with regard to test scores.
66. Recognizing unethical, illegal, or otherwise inappropriate assessment methods.
67. Recognizing unethical, illegal, or otherwise inappropriate uses of assessment information.

Copyright© 1994 by Zhicheng Zhang & Judith A. Burry-Stock.

APPENDIX B
Factor Structures of the Assessment Practices Inventory ($N = 297$)

Assessment Practices							*Self-Perceived Assessment Skills*							
Item	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*Item*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
13	.74						12	.75						
16	.73						14	.73						
14	.70						13	.71						
12	.68						15	.70						
4	.68						16	.68						
15	.63						4	.66						
17	.62						17	.64						
52	.61						32	.52						
18	.59						18	.51						
19	.57						5	.51						
51	.55						2	.50						
49	.52						3	.49						
32	.44						19	.47						
5	.42						52	.45						
3	.42						11	.39						
64	.36						1	.36						
58	.33						36		.72					
50	.30						35		.72					
2	.29						37		.69					
33		.72					34		.67					
34		.71					33		.65					
38		.68					38		.63					
35		.66					46		.61					
36		.65					47		.52					
46		.60					43		.52					
37		.52					9		.51					
43		.50					40		.49					
40		.50					25		.47					
25		.48					39		.47					
39		.46					8		.46					
9		.45					29			.77				
8		.40					28			.76				
47		.35					27			.67				
61			.56				30			.60				
67			.56				26			.59				
62			.54				24			.56				
59			.54				7			.54				
66			.53				31			.53				
41			.51				6			.45				
1			.50				10			.42				
11			.50				61				.65			
44			.49				62				.64			

(*continued*)

APPENDIX B (*Continued*)

Assessment Practices							*Self-Perceived Assessment Skills*							
Item	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*Item*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
42			.48				60				.57			
65			.47				63				.56			
48			.42				41				.49			
45			.41				65				.44			
63			.41				42				.42			
60			.36				59				.38			
29				.75			64				.36			
28				.72			56					.74		
27				.68			54					.72		
26				.56			55					.71		
30				.48			53					.64		
24				.48			57					.63		
7				.46			50					.30		
31				.36			45						.50	
6				.33			44						.49	
56					.76		22						.47	
55					.71		23						.47	
53					.64		49						.46	
54					.61		21						.44	
57					.47		58						.41	
21						.58	48						.41	
22						.55	51						.41	
23						.53	20						.38	
10						.44	67							.73
20						.44	66							.72
	F1	*F2*	*F3*	*F4*	*F5*	*F6*		*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
SS[a]	7.48	5.73	5.72	4.06	2.86	2.77	SS[a]	8.16	6.86	5.88	4.40	3.91	3.58	2.58
Var.[b]	11.2	8.6	8.5	6.1	4.3	4.1	Var.[b]	12.2	10.2	8.8	6.6	5.8	5.3	3.9

Note. For Assessment Practices: F1 = Factor 1 (Using Paper-Pencil Tests); F2 = Factor 2 (Standardized Testing, Test Revision, and Instructional Improvement); F3 = Factor 3 (Communicating Assessment Results, Ethics, and Grading); F4 = Factor 4 (Using Performance Assessment); F5 = Factor 5 (Nonachievement-Based Grading); F6 = Factor 6 (Ensuring Test Validity and Reliability).

For Self-Perceived Assessment Skills: F1 = Factor 1 (Perceived Skillfulness in Using Paper-Pencil Tests); F2 = Factor 2 (Perceived Skillfulness in Standardized Testing, Test Revision, and Instructional Improvement); F3 = Factor 3 (Perceived Skillfulness in Using Performance Assessment); F4 = Factor 4 (Perceived Skillfulness in Communicating Assessment Results); F5 = Factor 5 (Perceived Skillfulness in Nonachievement-Based Grading); F6 = Factor 6 (Perceived Skillfulness in Grading and Test Validity); F7 = Factor 7 (Perceived Skillfulness in Addressing Ethical Concerns).

Unless specified otherwise, the numbers represent factor loadings.

[a]sum of squared factor loadings for each factor. [b]percent of the variance explained by each factor after rotation.

Copyright © 2003 EBSCO Publishing