DOCUMENT RESUME

ED 408 333 TM 026 586

AUTHOR Zhang, Zhicheng; Burry-Stock, Judy
TITLE Assessment Practices Inventory: A Multivariate Analysis of Teachers' Perceived Assessment Competency.
PUB DATE Mar 97
NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Competence; *Educational Assessment; Educational Practices; Elementary Secondary Education; Evaluation Methods; Measurement Techniques; *Multivariate Analysis; Psychometrics; Teacher Attitudes; Teacher Education; *Teachers; *Teaching Experience; Test Reliability; Test Validity
IDENTIFIERS Composite Scores

ABSTRACT

A study was conducted to determine the psychometric properties and the subscales of a 67-item Assessment Practices Inventory (API) and to examine the effects of measurement training and teaching experience on teachers' perceived assessment competency. Data were collected from 311 teachers on the API. The reliability of the API was supported by a Cronbach alpha of 0.97. Construct validity of the API was examined using Rasch model and factor analyses. Based on the factor analysis, seven composite scores were formed on which a 2 x 3 multiple analysis of variance was conducted to examine the effects of measurement training and teaching experience on teachers' perceived competence in seven assessment categories. Multivariate interaction effects between measurement training and years of teaching were significant ($p<0.05$). Subsequent examination revealed significant multivariate simple effects of measurement training at 4 or more years of teaching in two factor-analyzed assessment categories ($p<0.01$). Followup comparisons between the means indicated that among the teachers who had taught 4 or more years, those with measurement training believed they were more skilled than those without measurement training in two main assessment categories ($p<0.001$; $p<0.05$). Implications for measurement training are discussed. (Contains 5 tables and 25 references.) (Author/SLD)

Assessment Practices Inventory: A Multivariate Analysis of Teachers' Perceived Assessment Competency

Zhicheng Zhang
Virginia Military Institute

Judy Burry-Stock
The University of Alabama

Paper presented at the annual meeting of
the National Council on Measurement in Education
Chicago, March 24-28, 1997

Abstract

The study was intended to (1) determine the psychometric properties and the subscales of a 67-item Assessment Practices Inventory (API) and (2) examine the effects of measurement training and teaching experience on teachers' perceived assessment competency. Data were collected from 311 teachers on the API. The reliability of the API was supported by a Cronbach alpha of .97. Construct validity of the API was examined using Rasch model and factor analyses. Based on the factor analysis, seven composite scores were formed on which a 2x3 MANOVA was conducted to examine the effects of measurement training and teaching experience on teachers' perceived competency in seven assessment categories. Multivariate interaction effects between measurement training and years of teaching were significant ($p < .05$). Subsequent examination revealed significant multivariate simple effects of measurement training at four or more years of teaching in two factor-analyzed assessment categories ($p < .01$). Follow up comparisons between the means indicated that among the teachers who had taught four or more years, those with measurement training believed they were more skilled than those without measurement training in two main assessment categories ($p < .001$; $p < .05$). Implication for measurement training is discussed.

Review of Related Literature

Measurement training for classroom teachers has received increasing attention in recent years. Three related themes can be found in the literature on classroom assessment: delineating the content domain for measurement training for teachers, identifying problem areas in teacher assessment skills, and investigating teacher beliefs about their knowledge of testing in relation to measurement training.

Numerous researchers and organizations have specified the content domain in which teachers need to develop assessment skills (Airasian, 1994; American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990; Burry-Stock, 1995; Carey, 1994; Schafer, 1991; Stiggins, 1992b; Zhang & Nejad, 1993). Among the commonly discussed skills are choosing appropriate assessment methods; developing paper-pencil tests; developing performance measures; administering and scoring tests; interpreting standardized test results; evaluating and improving assessment instruments; using assessment in decision making; grading; communicating assessment results; and ethics in assessment.

Teachers are found to be inadequately prepared for classroom assessment. Problems are particularly prominent in using performance measures, interpreting standardized test results, and grading. When using performance measures, for example, many teachers did not define levels of performance or plan scoring procedures in advance, nor did they record their scoring during assessment (Stiggins, 1992a). For standardized tests, teachers reported engaging in inappropriate practices of teaching test items, increasing time limits, giving hints, and changing students' answers (Hall & Kleine, 1992; Nolen, Haladyna & Haas, 1992). Most teachers also had trouble understanding and interpreting percentile ranks, grade equivalent scores, and percentile bands (Hills, 1991; Impara, Divine, Bruce, Liverman, & Gay, 1991). When assigning grades, many teachers incorporated nonachievement factors of efforts, attitude, and motivation into grades (Griswold, 1993; Hills, 1991; Jongsma, 1991; Stiggins, Frisbie, & Griswold,1989) and they often did not use weights to distinguish between formative and summative data.

Despite the problems mentioned above, most teachers believed they had adequate knowledge of testing (Gullickson, 1984) and they attributed their knowledge of testing and measurement more to experience than to university course work (Gullickson, 1984; Wise,

Lukin, & Roos, 1991). More teachers with more measurement training rated university course work as the second important factor that affects their knowledge of testing than teachers with less measurement training. These research suggests that experience and measurement training are perceived by teachers to be two major factors affecting their knowledge of testing. However, little has been done to determine in what specific assessment areas experience and/or measurement training help teachers advance to a higher skill level. The purpose of the present study was to extend the findings of Gullickson and Wise et. al. by investigating simultaneously the effects of experience and measurement training on teacher beliefs about their assessment competency. Specifically, the study was designed to (1) determine the psychometric properties and the subscales of the Assessment Practices Inventory (API) and (2) investigate how experience and measurement training affect teachers' perceived assessment competency in the assessment areas represented by the subscales of the API. In this study, experience was operationalized by years of teaching. The investigation of teachers' perceived assessment competency was conducted within the broad framework of classroom assessment skills specified in the literature.

Method

Instrument

The instrument used in the study was the API (Zhang and Burry-Stock, 1994). The instrument consists of 67 items each of which described an assessment practice. The items cover the broad spectrum of classroom assessment and reflect the spirit of the current literature on classroom assessment. For each item, the respondents were asked to report their perceived assessment competency on a 5-point scale with 1 meaning "NOT AT ALL SKILLED" and 5 meaning "HIGHLY SKILLED." Information was also collected on demographic variables concerning the number of years the teachers had taught and the number of measurement courses they had taken. Sample items are presented in Table 1.

Insert Table 1 About Here

The content validity of the instrument was built into the construction process by developing the items according to the content domain of classroom assessment specified in the literature. The construct validity of the API was examined using factor analysis and the

Rasch model with the computer program BIGSTEPS (Linacre & Wright, 1994). The distribution of item logits from -.89 to 1.31 provided evidence that, to a degree, the items defined the theoretical construct of "perceived skill level" of classroom assessment (Wright & Stone, 1979; Zhang, 1995). The reliability of the API was supported by a Cronbach alpha of .97 and item-to-total correlations all above .37. The standard error of measurement for the total score was 7.7.

Sample and Procedure

Surveys were sent to 845 teachers in two Alabama school districts. One school district was predominantly rural and the other was predominantly urban. The numbers of elementary, middle school/junior high, and high schools participating in the study were 6, 4, and 6, respectively. A vocational school was also included in the study. This was done to ensure a balanced representation of the teachers from different grade levels. The instrument, together with a cover letter and computer scanable answer sheet, was distributed to the teachers by their school principal at a faculty meeting. Those who voluntarily responded to the survey returned the completed answer sheets to the school secretary.

Three hundred and eleven completed surveys were collected. The teachers responding to the survey were predominantly white (89%) and female (77.4%). The distribution of respondents, by level of school taught, was as follows: elementary school, 34%; middle school/junior high, 23%; high school, 30%, respectively. Comprehensive and other types of schools made up the remaining 13%. Forty percent of these teachers obtained a bachelor's degree, 56% had a Master's degree. About 82% of the teachers had had at least one measurement course. The average number of years in teaching was 10.9. The breakdown by years of teaching was as follows: one or less than one year, 14%; two to three years, 12%; four or more years, 73%. About 1% of the teachers did not supply information on years of teaching.

Results

Preliminary Analysis

To identify the underlying dimensions of the API, a principal factor analysis was conducted with principal axis method of extraction and a varimax orthogonal rotation. Seven factors were retained. The eigenvalues of the seven factors were 8.20, 6.77, 5.57, 4.58, 3.96, 2.97, and 2.73, respectively. The seven factors accounted for 51.91% of the

variance. The emergence of the seven factors suggests that the 67-item API measures seven major assessment categories as summarized here and in Table 3:

1. Develop and administer paper-pencil tests, choose tests,
2. Interpret standardized test results, calculate test statistics, use assessment results in decision making,
3. Develop and use performance assessment, informal assessment,
4. Communicate test results,
5. Non-Achievement based grading,
6. Ethics in assessment, and
7. Grading.

Table 2 shows rank ordered factor loadings of individual items for each factor. Only high loadings (greater than .31) selected by the computer program are presented. The number of items, the means, the standard deviations, and Cronbach alpha reliability coefficients for the subscales of the API are presented in Table 3.

Insert Tables 2 and 3 About Here

A 2x3 MANOVA

Based on the factor analysis, a composite score was calculated for each identified subscale of the API. The composite scores served as the dependent variables. The two independent variables were measurement training (no training, at least one measurement course) and years of teaching (one or less year of teaching, two-three years of teaching, four or more years of teaching). A 2x3 MANOVA was then conducted using general linear model in SAS (SAS, 1995)

The 2x3 MANOVA revealed significant multivariate interaction effects between measurement training and years of teaching on teachers' perceived assessment skills in seven major categories ($F = 1.61$, $p < .05$). Subsequent examinations indicated significant multivariate simple effects of training at four or more years of teaching ($F = 3.3$, $p < .01$). The multivariate interaction and simple effects are presented in Table 4.

Follow-up examinations of the univariate simple effects of measurement training at four or more years of teaching in the seven assessment categories revealed significant

univariate simple effects in the category of interpreting standardized test results, calculating test statistics, and using assessment results in decision making (F=12.74, p <.001). Significant univariate simple effects were also found in the category of developing performance assessment and using informal observation (F=4.26, p <.05). The univariate simple effects are presented in Table 5.

Insert Tables 4 and 5 About Here

Follow up comparisons between the means indicated that among the teachers who had taught four or more years, those with measurement training scored significantly higher than those without measurement training in interpreting standardized test results, calculating test statistics, and using assessment results in decision making (46.93 versus 39.46). Among the teachers who had taught four or more years, those with measurement training scored significantly higher than those without measurement training in developing performance assessment and using informal observation (40.86 versus 37.90).

Discussion and Conclusion

Based on the factor analysis, seven factors were identified for the 67-item API. Accounting for 51.91 % of the variance, the seven factors represent the major assessment categories the API is intended to measure. The subscales supported by the factor structure were then used as the assessment categories in which we examined the effects of measurement training and teaching experience on teachers' perceived assessment competency.

One noticeable feature about the API was that the items designed to measure test construction and test use loaded on two factors: the factor of (mainly) paper-pencil tests and the factor of (mainly) performance assessment. This result suggests paper-pencil tests and performance assessment are related but different assessment categories. The two assessment methods require different techniques and skills. Teachers considered themselves less skilled in using performance assessment (Zhang, 1995) and there is a growing concern in the assessment community about validity, reliability, and authenticity of performance assessment (Baron, 1991; Brandt, 1992; Dunbar, Koretz, & Hoover, 1991). The emergence of paper-pencil tests and performance assessment as two separate subscales in the present study

enables the researchers to examine how measurement training and teaching experience affect teachers' perceived competency in two major areas of test construction.

The multivariate analysis of the data suggested that the teachers with measurement training and at least four years of teaching experience believed they were more skilled than the teachers with similar teaching experience but without measurement training in interpreting standardized test results, calculating test statistics, and using assessment results in decision making. The teachers with measurement training and at least four years of teaching also perceived themselves to be more skilled in using performance assessment and informal observation than the teachers with similar teaching experience but without measurement training.

Interpreting standardized test results is an assessment area in which teachers were found to have problems (Hill, 1991). Teachers had trouble interpreting a percentile band performance profile even with the help of interpretive information (Impara, Divine, Bruce, Liverman, & Gay, 1991). Zhang's research (1995) using Rasch model analysis suggested that interpreting standardized test results, conducting classroom assessment, and using assessment results in decision making were perceived by teachers to be the most difficult assessment category. Yet it is in this assessment category that teachers with both measurement training and teaching experience reported having a higher degree of perceived competency. One possible reason for this is that interpreting standardized test results and calculating test statistics involve technical skills that are often taught in a university measurement course while the skills of using assessment results in decision making may require both testing knowledge and teaching experience.

The present study also found that teachers with measurement training and at least four years of teaching experience perceived themselves to be more skilled than the teachers with similar teaching experience but with no measurement training in performance assessment and informal observation. Previous research has suggested that teachers considered performance assessment to be the third most difficult assessment category, even more so than paper-pencil tests (Zhang, 1995). Teachers in general are less proficient in performance assessment and they often fail to follow the recommended practices in constructing and using performance instruments and communicating assessment criteria (Stiggins, 1992a). However, it is in this assessment category that teachers with measurement training and teaching

experience believed they were more skilled than those with similar teaching experience but without measurement training . These two findings suggest the value of measurement training and teaching experience.

The present study strongly supports the need for measurement training whether it be done through university course work or district inservice programs. In addition to a broad coverage of assessment techniques, special attention should be directed to interpreting standardized test results, using assessment results in decision making, calculating test statistics, developing and using performance assessment, and using informal observation. It is crucial that teachers acquire adequate measurement and testing skills through appropriate training and use those skills in assessing student learning and achievement.

In conclusion, the research findings provide evidence concerning the effects of measurement training and teaching experience. These results are self-report. Designed to gather perceived information about assessment practices, the API can be very useful as a diagnostic tool for determining staff needs. Future investigations should focus on exploring how teachers use assessment practices to evaluate student learning and achievement. An interesting study would be to interview teachers and observe first hand how they actually conduct assessment in the classroom.

Since the self-report inventory was used only with 311 teachers mainly from two local school districts in a southeastern state, the present research findings should be interpreted with caution. The replication of the study with a larger sample is desired to confirm the present research findings.

References

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

Airasian, P. W. (1994). Classroom assessment. McGraw-Hill, Inc.

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

Brandt, R. (1992). On performance assessment: A conversation with Grant Wiggins. Educational Leadership, 49(8), 35-37.

Burry-Stock, J. A. (Ed.). (1995). BER 450 Handbook. Tuscaloosa, AL: The University of Alabama.

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

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.

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

Gullickson, 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.

Linacre, J. M. & Wright, B. D. (1994). A user's guide to BIGSTEPS. MESA Press: Chicago.

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.

SAS, (1995). SAS Institute, Inc.

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. Educational Measurement: Issues and Practice, 10(1), 3-6.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. Educational Measurement: Issues and Practices, 8(2), 5-14.

Stiggins, R. J. (1992a). In teachers' hands: Investigating the practices of classroom assessment. Albany: State University of New York Press.

Stiggins, R. J. (1992b). High quality classroom assessment: What does it really mean? Educational Measurement: Issues and Practice, 11(2), 35-39.

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.

Wright, B. D. & Stone, M. H. (1979). Best test design. MESA Press: Chicago.

Zhang, Z. (1995). Investigating Teachers' Perceived Assessment Practices and Assessment Competencies on the Assessment Practices Inventory (API). Unpublished doctoral dissertation. University of Alabama, AL: Tuscaloosa.

Zhang, Z. & Iran-Nejad, A. (1993, November). A thematic approach to teaching tests and measurement . Paper presented at the annul meeting of Mid-South Educational Research Association. New Orleans, LA.

Zhang, Z. & Burry-Stock, J. A. (1994). Assessment practices inventory. Tuscaloosa, AL: The University of Alabama.

Table 1
Sample Items of the API

Respond to each statement using the following scale:

NOT AT ALL SKILLED 1 / 2 / 3 / 4 / 5 HIGHLY SKILLED

1. Developing assessments based on clearly defined course objectives.
2. Ensuring adequate content sampling for a test.
3. Constructing a model answer for scoring essay questions.
4. Defining a rating scale for performance assessment in advance.
5. Following required procedures (time limits, no hint, no interpretation) when administering standardized tests.
6. Interpreting percentile bands to students and parents.
7. Weighing differently projects, exams, homework, etc. when assigning semester grades.
8. Using assessment results when making decisions about individual students (e.g., placement, graduation).
9. Communicating assessment results to parents.
10. Protecting students' confidentiality with regard to test scores.

Table 2
Factor Loadings of the API: A Seven-Factor Solution With a Varimax Rotation (n=311)

Factor 1		Factor 2		Factor 3		Factor 4		Factor 5		Factor 6		Factor 7	
Item	Loading	Item	Loading	Item	Loading	Item	Loading	Item	Loading	Item	Loading	Item	Loading
14	73	33	70	29	76	60	57	56	75	67	69	45	46
12	72	34	69	28	74	62	52	54	73	66	66	44	44
15	72	35	69	27	67	63	49	55	71			23	39
13	69	36	68	30	62	65	47	57	63			48	39
16	69	38	65	26	56	42	44	53	63				
4	65	37	64	24	55	59	43	50	31				
17	61	46	57	31	53	51	43						
2	52	39	54	7	48	64	40						
5	52	43	54	21	45	58	34						
18	51	40	53	6	43								
49	50	9	48	22	43								
32	50	47	47	61	57								
3	49	25	46										
19	48	8	46										
10	43	41	45										
20	41												
52	41												
11	41												
1	40												
*	8.20		6.77		5.57		4.58		3.96		2.97		2.73
**	12.23		10.10		8.30		6.83		5.91		4.43		4.10

* sum of squared factor loadings
** percent of variance explained by each factor

Note: Factor loadings are multiplied by 100 to remove decimals.

Table 3
Descriptive Statistics for the API Subscales Emerged from Factor Analysis n=311

Subscales	# of items	Mean	SD	Cronbach Alpha
Develop and use Paper-pencil tests, choose tests	19	72.70	13.24	.93
Interpret standardized test results, calculate test statistics, use assessment results in decision making,	15	45.59	11.95	.90
Develop and use performance assessment, informal assessment	12	44.44	8.71	.89
Communicate test results	9	34.90	6.28	.84
Non-achievement based grading	6	20.08	5.78	.85
Ethics in assessment	2	7.89	2.19	.87
Grading	4	14.95	3.48	.79

Table 4
A 2x3 MANOVA: (1) Multivariate Interaction Effects of Measurement Training by Years of Teaching and (2) Multivariate Simple Effects of Measurement Training at Four or More Years of Teaching n=311

	Wilks' Lambda	F Value	p Value
(1) Multivariate Interaction Effects	.82	1.61	.0146*
(2) Multivariate Simple Effects	.92	3.3	.0022**

significant at alpha=.05 ** significant at alpha=.01

Table 5
A 2x3 MANOVA: Univariate Simple Effects of Measurement Training at Four or More Years of Teaching
n=311

API Subscales	F Value	p Value
1. Develop and administer paper-pencil tests, choose tests	1.52	.2187
2. Interpret standardized test results, calculate test statistics use assessment results in decision making	12.74	.0004 **
3. Develop and use performance assessment, informal assessment	4.26	.0399 *
4. Communicate test results	2.76	.0977
5. Non-Achievement based grading	.19	.6672
6. Ethics in assessment	2.58	.1091
7. Grading	.22	.6391

* significant at alpha=.05 ** significant at alpha=.01