Post Hoc Analysis of Test Items Written by Technology Education Teachers

W. J. Haynie, III

Technology education teachers frequently author their own tests. The effectiveness of tests depends upon many factors, however, it is clear that the quality of each individual item is of great importance. This study sought to determine the quality of teacher-authored test items in terms of nine rating factors.

Background

Most testing in schools employs teacher-made tests (Haynie, 1983, 1990, 1991; Herman & Dorr-Bremme, 1982; Mehrens & Lehmann, 1987; Newman & Stallings, 1982). Despite this dependance upon teacher-made tests, Stiggins, Conklin, and Bridgeford (1986) point out that "nearly all major studies of testing in the schools have focused on the role of standardized tests" (p. 5).

Research concerning teacher-constructed tests has found that teachers lack understanding of measurement (Fleming & Chambers, 1983; Gullickson & Ellwein, 1985; Mehrens & Lehmann, 1987; Stiggins & Bridgeford, 1985). Research has shown that teachers lack sufficient training in test development, fail to analyze tests, do not establish reliability or validity, do not use a test blueprint, weight all content equally, rarely test above the basic knowledge level, and use tests with grammatical and spelling errors (Burdin, 1982; Carter, 1984; Gullickson, 1982; Gullickson & Ellwein, 1985; Hills, 1991). Technically their tests are simplistic and depend upon short answer, true-false, and other easily prepared items. Their multiple-choice items often have serious flaws--especially in distractors (Haynie, 1990; Mehrens & Lehmann, 1984, 1987; Newman & Stallings, 1982).

A few investigations have studied the value of tests as aids to learning subject content (Haynie, 1987, 1990, 1991; Nungester & Duchastel, 1982). Time on-task has been shown to be very important in many studies (Jackson, 1987; Salmon, 1982; Seifert & Beck, 1984). Taking a test is a time on-task learning activity. Works which studied testing versus similar on-task time spent in structured review of the material covered in class have had mixed results, but testing appears to be at least as effective as reviews in promotion of learning

W.J. Haynie, III is Associate Professor, Department of Occupational Education, North Carolina State University, Raleigh, NC.

(Haynie, 1990; Nungester & Duchastel, 1982). Research is lacking on the quality of tests and test items written by technology education teachers.

Purpose

The purpose of this investigation was to study the quality of technology education test items written by teachers. Face validity, clarity, accuracy in identifying taxonometric level, and rates of spelling and punctuation errors were some of the determinants of quality assessed. Additionally, data were collected concerning teachers' experience levels, highest degree held, and sources of training in test construction. The following research questions were addressed in this study:

1. What types of errors are common in test items?
2. Do the error rate or types of errors in teacher constructed test items vary with demographic factors?
3. Do teachers understand how to match test items to curriculum content and taxonometric level?

Methodology

Source of Data

Between April 23, 1988 and January 8, 1990, a team of 15 technology education teachers worked to develop test items for a computerized test item bank for the North Carolina State Department of Public Instruction (SDPI). The work was completed under two projects funded by SDPI and directed by DeLuca and Haynie (1989, 1990) at North Carolina State University. The data for this study came from the items developed in those projects.

Test Item Authors

The teachers were selected on recommendation of supervisors, SDPI consultants, or teacher educators. All were recognized as leaders among their peers and most had been nominated for teacher of the year or program of the year commendation. They were all active in the North Carolina Technology Education Association and supported the transition to the new curriculum. Table 1 displays demographic data concerning the test item authors.

Table 1
Profile of Authors' Demographic Factors

Author	Years of Teaching Experience	Highest Degree	Undergraduate Test & Measure Courses	Graduate Test & Measure Courses
1	9	B.S.	0	0
2	5	B.S.	1	0
3	23	B.S.	0	0
4	4	B.S.	0	1
5	5	B.S.	0	1
6	23	M.Ed.	0	1
7	19	M.Ed.	0	1
8	17	M.Ed. + 2 yrs.	0	2
9	25	M.Ed.	0	0
10	5	M.Ed.	0	0
11	7	M.Ed.	0	0
12	7	B.S.	0	0
13	7	M.Ed.	0	0
14	15	B.S.	1	0
15	5	B.S.	1	1

Training of Authors

Teachers came to the university campus for a workshop on April 23, 1988. Project directors oriented teachers to the computerized test bank, reviewed the revised technology education curriculum, and explained how to develop good test items. A 13 page instructional packet was also given to each author. It should be noted that the training session and instructional packet may confound attempts to generalize these findings.

The authors were required to develop and properly code six items which were submitted for approval and corrective feedback before they were allowed to proceed. The teachers who authored the items were paid an honorarium for their services.

Editing and Coding of Items

Each item was prepared on a separate sheet of paper with a coding sheet attached and completed by the teacher. The coding sheet identified the author, the specific objective tested, the taxonometric level, and information for the computerized system. The project directors edited the items with contrasting colored felt tip pens on the teachers' original forms.

Design of this Study

The data for this investigation were the editing markings on the original test items submitted by the teachers. Scores for 9 scales of information were recorded for analysis. Each of the scales was established so that a low score would be optimal. The scales were Spelling Errors (SE), Punctuation Errors (PE), Distractors (D), Key (K), Usability (U), Validity (V), Stem Clarity (SC), Taxonomy (TX), and an overall Quality (Q) rating. After all of the ratings were completed, the General Linear Models (GLM) procedure was used for F testing and the LSD procedure was used when t-tests were appropriate.

Findings

Spelling Errors (SE)

The frequency and percentage of scores for the 993 items on the nine ratings, and mean scores of each factor, are shown in Table 2. An item's SE rating indicates how many words were misspelled in the item. There were 98 items (10%) which had one or more spelling errors. Spelling errors are detrimental to good teaching and testing. However the literature shows that this problem is common to other disciplines.

Table 2
Ratings of Test Item Quality

Rating Category	Score	Frequency of Items With Each Score	% of Items/ Score	Mean Item Score	SD
Spelling Errors (SE)	0	895	90.1		
	1	76	7.7		
	2	11	1.1		
	3	6	0.6		
	4	3	0.3		
	5	1	0.1		
	6	1	0.1		
SE Totals	---	993	100%	0.14	0.52
Punctuation Errors (PE)	0	735	74.0		
	1	220	22.2		
	2	25	2.5		
	3	4	0.4		
	4	1	0.1		
	5	8	0.8		
PE Totals	---	993	100%	0.33	0.68

Table 2 (cont.)

Distractors (D)		0	447	45.0		
		1	398	40.1		
		2	95	9.6		
		3	30	3.0		
		4	9	0.9		
		5	14	1.4		
	D Totals	---	993	100%	0.79	0.96
Key (K)		0	889	89.5		
		2	104	10.5		
	K Totals	---	993	100%	0.21	0.61
Usability (U)		0	249	25.1		
		1	265	26.7		
		2	159	16.0		
		3	131	13.2		
		4	74	7.5		
		5	50	5.0		
		6	21	2.1		
		7	11	1.1		
		8	16	1.6		
		9	17	1.7		
	U Totals	---	993	100%	2.02	2.04
Stem Clarity (SC)		0	602	60.6		
		1	352	35.4		
		2	39	3.9		
	SC Totals	---	993	100%	0.43	0.57
Taxonomy (TX)		0	835	84.1		
		1	124	12.5		
		2	34	3.4		
	TX Totals	---	993	100%	0.19	0.47
Quality (Q)		0	208	20.9		
		1	235	23.7		
		2	200	20.1		
		3	129	13.0		
		4	74	7.5		
		5	58	5.8		
		6	42	4.2		
		7	17	1.7		
		8	10	1.0		
		9	12	1.2		
		10	2	0.2		
		11	3	0.3		
		12	1	0.1		
		13	1	0.1		
		14	1	0.1		
		15	0	---		
		16	0	---		
		17	1	0.1		
	Q Totals	----	993	100%	2.28	2.20

Note. There were 993 items.

The authors were compared on each of the scales to determine whether they differed significantly and to see if similar or dissimilar errors were made by different authors. On the spelling errors factor authors were found to differ significantly: $F(14, 978) = 11.99$, $p<.0001$. Follow-up analysis with the LSD procedure showed that 5 authors had significantly fewer spelling errors and 3 authors had more than the average number of errors in spelling (Table 3). Two of the authors with numerous spelling errors also had other defects and were rated significantly worse in the overall Quality (Q) rating (authors 1 and 9). However, only 1 of the authors with a significantly low rate of spelling errors was rated favorably in the Quality rating, so spelling accuracy alone is insufficient to identify good test item writing ability.

Table 3
Means of each Author on the 9 Rating Categories

	N	Per Item Means								
Author	Items	SE	PE	D	K	U	V	SC	TX	Q
1	92	0.29 **	0.37	1.29 **	0.68 **	2.95 **	0.09 *	0.53	0.24	3.51 **
2	102	0.01 *	0.17 *	0.59	0.12	1.34	0.16	0.38	0.11	1.54
3	32	0.21	0.41	1.16 **	0.44	2.88 **	0.28 **	0.47	0.59 **	3.56 **
4	103	0.17	0.39	1.28 **	0.33	2.76 **	0.16	0.49	0.17	2.98
5	100	0.17	0.39	0.94	0.24	3.01 **	0.22	0.67 **	0.29 **	0.92
6	56	0.11	0.38	1.14 **	0.32	2.25	0.32 **	0.43	0.34 **	3.04
7	62	0.26 **	0.24	0.55	0.26	1.77	0.13	0.39	0.35 **	2.18
8	104	0.07 *	0.22	0.71	0.17	1.70	0.38 **	0.38	0.19	2.13
9	42	0.43 **	0.83 **	1.21 **	0.09	3.21 **	0.26 **	0.79 **	0.29 **	3.90 **
10	50	0.04 *	0.98 **	0.16 *	0.00 *	1.46	0.00 *	0.28 *	0.04 *	1.50
11	46	0.00 *	0.28	0.74	0.00 *	1.85	0.13	0.35	0.09 *	1.59
12	28	0.21	0.07 *	0.39	0.00 *	1.04 *	0.11	0.29	0.18	1.25 *

Table 3 (cont.)

13	82	0.06 *	0.01 *	0.14 *	0.02 *	0.71 *	0.30 **	0.23 *	0.07 *	0.85 *
14	48	0.13	0.31	0.29 *	0.00 *	1.19 *	0.04 *	0.42	0.08 *	1.27 *
15	46	0.09	0.17 *	0.87	0.04 *	1.54	0.00 *	0.26 *	0.00 *	1.43
Grand Means	---	0.14	0.33	0.79	0.21	2.02	0.19	0.43	0.19	2.28

Note. There were 993 items.
* Significantly low (better), $p<.05$.
** Significantly high (worse), $p<.05$.

Years of teaching experience and other demographic data were presented in Table 1. Teachers were divided into two groups of experience level: fewer than 8 years experience (8 teachers who authored 557 items), and more than 8 years experience (7 authors, 436 items). On the Spelling Errors factor these groups were compared and there was a significant finding of $F(1, 991) = 10.48$, $p<.0012$. Follow-up analysis by the LSD procedure showed that the less experienced teachers had significantly fewer spelling errors. None of the other demographic variables were found to differ significantly on the rate of spelling errors.

Punctuation Errors (PE)

The PE rating (Table 2) was the total number of punctuation errors. The most frequent errors were omission of punctuation at the end of the stem or use of the wrong punctuation there. Frequently statements were ended with question marks or stems which should have ended with a colon were left with no punctuation. This score may be inflated spuriously by those unique errors which may not have been made in normal prose writing by the same teachers. Among the 15 authors, a significant difference was found in the PE category: $F(14, 978) = 8.12$, $p<.0001$ (Table 3). No significant differences were found among any demographic variables on the rate of punctuation errors.

Distractors (D)

Errors in distractors other than spelling or punctuation were summed in the Distractors (D) category (Table 2). Frequently these errors either eliminated distractors or targeted the correct answer due to incompatibility between the stem and the alternatives because of lack of agreement in: singular-plural, introductory article, tense, or in one case even gender.

A significant finding of $F(14, 978) = 13.37$, $p<.0001$ was attained and follow-up by LSD showed that 3 authors (10, 13, and 14) had significantly

lower error rates. Two of those 3 authors who had superior distractors were also among the best in the overall Quality rating. All three of the authors who rated poorest in the overall Quality rating, also rated significantly worse in this Distractors category. Apparently this is one aspect of test writing which needs to be stressed to teachers.

All 4 of the demographic variables studied were found to be significantly related to errors in distractors: Years of experience, $F(1, 991) = 10.55$, $p<.0012$, the less experienced teachers authored superior distractors; Highest degree held, $F(1, 991) = 23.21$, $p<.0001$, those with graduate degrees wrote better distractors; Undergraduate courses, $F(1, 991) = 11.46$, $p<.0007$, those who had taken an undergraduate testing and measurement course prepared better distractors; and Graduate courses, $F(1, 991) = 13.23$, $p<.0003$, graduate courses also appeared beneficial.

Key (K)

The Key (K) rating simply indicates whether the answer marked in the teacher's original version of the item was indeed correct. Since incorrect keying was considered a more damaging error than a misspelled word or other common error, a rating of 2 was given for incorrectly keyed items. This resulted in greater increase of the summation categories (Usability and Quality) due to incorrect keying than for other types of errors. Regrettably, 10.5% of the items were keyed incorrectly (Table 2).

The authors differed significantly in the Key rating: $F(14, 978) = 8.01$, $p<.0001$. Table 3 shows the teachers' means and the results of LSD comparisons. Six authors keyed their items more accurately than others and one teacher was very inaccurate in keying. Teachers with less than eight years of experience keyed more accurately than more experienced teachers, $F(1, 991) = 19.82$, $p<.0001$; and teachers with graduate degrees also more accurately keyed their items, $F(1, 991) = 12.90$, $p<.0003$.

Usability (U)

The Usability (U) rating was found by counting all proofreading and editing marks of all types on the teachers' original forms--thus it included the sum of all the above categories plus other errors and defects not included in them. An example of an error which would not be counted in the first four ratings but would be included here is an item which begins with a blank. Such an item would have a U rating which equalled the sum of all SE, PE, D, and K ratings plus 1.

The teachers did differ significantly when compared on the Usability of their items: $F(14, 978) = 11.99$, $p<.0001$. Comparisons via LSD found that three teachers developed items with superior usability and five teachers authored significantly less usable items (Table 3). The teachers with fewer than eight years of experience developed more usable test items according to this rating: $F(1, 991) = 7.47$, $p<.0064$. Teachers with graduate degrees wrote more useful items, $F(1, 991) = 16.42$, $p<.0001$, and both undergraduate and graduate

testing and measurement courses appeared to be effective in helping teachers develop usable items: Undergraduate courses, $F(1, 991) = 26.68$, $p<.0001$; and Graduate courses, $F(1, 991) = 12.05$, $p<.0005$.

Validity (V)

Items were carefully read and compared to the objectives they were intended to test. A Validity (V) rating of 0 indicated the item clearly possessed face validity. An item which was obviously off the subject was rated 2 and items which tested information immediately adjacent to the intended information were rated 1 to indicate that validity was questionable.

The authors differed significantly in how valid their items appeared to be: $F(14, 978) = 3.99$, $p<.0001$. It is noteworthy that the Validity rating did not necessarily correspond to others in the study. One of the authors (number 1) who rated significantly better in terms of validity was one of the worst rated authors in five other categories. Likewise, one other author (number 13) who rated superior in eight other categories (including Q) was significantly worse in the Validity category.

The findings related to the demographic variables were: Less experienced teachers wrote more valid items, $F(1, 991) = 4.32$, $p<.038$; teachers with only Bachelor's degrees wrote more valid items than those with graduate degrees, $F(1, 991) = 11.47$, $p<0007$; teachers who had experienced undergraduate test and measurement courses submitted more valid items, $F(1, 991) = 9.29$, $p<.0024$; and graduate courses also helped teachers write more valid items, $F(1, 991) = 10.01$. $p<.0018$.

Stem Clarity (SC)

Stem Clarity (SC) was a subjective rating indicating how clearly understandable the stem appeared. If the item's stem seemed clear enough to lead knowledgeable students to the correct response, regardless of other types of errors (SE, PE, D, K, U, or V ratings), then that item was rated 0 in the SC category. Items which were confusing to read with no clear purpose set forth in the stem were rated 2. Items which would likely work but had some element of confusion were rated 1. Table 2 shows that most items were judged to be reasonably clear in intention.

The finding of $F(14, 978) = 4.57$, $p<.0001$ documents that teachers did vary in their ability to write clear item stems. It would seem reasonable to assume that authors who made many spelling and punctuation errors would also have difficulty wording their stems clearly. This, however, was not true in these findings. Of the demographic factors investigated, only highest degree held was related to the ability to prepare clearly worded stems: $F(1, 991) = 6.34$, $p<.0120$, teachers with graduate degrees developed superior items in terms of stem clarity.

Taxonomy (TX)

The Taxonomy (TX) rating indicates the extent to which teachers accurately identified the taxonometric level of the cognitive domain for each item. Teachers prepared items to match specific objectives and then coded them. The codes used were derived from the first three levels of Bloom's Taxonomy: 1 indicated simple knowledge, 2 indicated comprehension, and 3 indicated application or higher levels of learning.

Of the 993 items prepared for the test item bank, the authors indicated that they felt 559 (56%) operated at level 1 (knowledge), 379 (38%) operated at level 2 (comprehension), and only 55 (5.5%) operated at level 3 (application or above). The rating in the TX category assigned for this study indicates how well, in the researcher's judgement, the item authors had accurately identified the proper taxonometric level. This was done after reading the objective to be tested by each item and then carefully reading the item to see if it operated at the level indicated by the teacher. A rating of 0 in the TX category indicates that the item appeared to be accurately coded by the teacher. A rating of 2 indicated that there was a clear mismatch between the level at which the teacher desired the item to function and the level at which the researcher judged the item would actually operate. Ratings of 1 in the TX category indicate that the researcher felt the author's coding was questionable.

Table 2 shows that 84% (835) of the items had been correctly coded for taxonometric level. Teachers did vary significantly in their ability to code items according to taxonomy: $F(14, 978) = 5.20$, $p<.0001$. All teachers who rated poor in this rating also had poor ratings in at least one other category, most rated poor in at least two others. Teachers who rated superior in the TX rating also rated superior in at least two other ratings. Teachers with less than 8 years of experience were significantly more accurate in coding by taxonomy than the more experienced teachers, $F(1, 991) = 21.08$, $p<.0001$. Undergraduate test and measurement courses, $F(1, 991) = 9.29$, $p<.0024$, appeared to be helpful in enabling teachers to identify the correct taxonometric level of test items, however, graduate courses were not found to be a significant factor here, $F(1, 991) = 2.65$, $p<.0711$.

Quality (Q)

The overall Quality of the test items was summarized in the Q rating. The Q rating was found by summing all of the other ratings except Usability (U), which was already a partial summation. The Q ratings (Table 2) range from 0 (an item judged to need no editing of any sort and believed to operate exactly as the submitting author had intended) to a high value of 17.

A finding of $F(14, 978) = 14.79$, $p<.0001$, shows that teachers differed in Q ratings (see Table 3). All of the teachers who differed significantly in the Q rating had also differed in several other categories. Experienced teachers prepared items with poorer overall quality than inexperienced teachers: $F(1. 991) = 20.67$, $p<.0001$. Teachers with graduate degrees produced items identified to have better quality: $F(1, 991) = 13.44$, $p<.0003$. Undergraduate test

and measurement courses helped teachers develop higher quality items, $F(1, 991) = 35.45$, $p<.0001$, and so did graduate courses, $F(1, 991) = 11.14$, $p<.0009$.

Discussion

Though the sample included only 15 teachers, the findings presented in this study suggest that technology education teachers have some of the same difficulties in developing useful test items that teachers in other disciplines face. Despite the fact that these carefully selected teachers were given special training to improve their items, less than 21% of the items they prepared were flawless. Earlier works identified spelling, punctuation, grammar, clarity, validity, reliability, taxonometric level, problems in distractors, and other mechanical factors to be problem areas in teacher-made tests. Six of these problems were investigated in this study. Additionally, errors in keying items, a general overall quality assessment, and preparation of technology education teachers to write test items were factors considered by this study.

It was demonstrated that teachers differed significantly in their ability to prepare good test items, and that undergraduate and graduate courses in testing and measurement, though they appear to be helpful in many ways, are not taken by all teachers. These courses improved teachers' ability in developing distractors, and preparing valid and useful items. Undergraduate courses were also shown to help teachers identify the proper taxonometric level of their items.

Teachers with graduate degrees developed items which were superior in 5 of the ratings in this study: distractors, keying of items, usability, stem clarity, and overall quality. However, teachers who had only Bachelor's degrees were significantly better in developing items judged to have good face validity.

Teachers with fewer than 8 years of experience developed items with better overall quality (Q rating) than those who had more experience. The less experienced teachers significantly outperformed their more experienced peers on 7 of the quality factors studied: spelling, distractors, key accuracy, usability, validity, taxonomy, and overall quality. These findings were unanticipated and could possibly be explained by any of several competing theories. Perhaps teachers who have been in the profession longer than 8 years have begun to burn out and have less time or patience to devote to extra assignments such as the test item development projects in which they participated. Alternatively, it could simply be true that teachers who earned their degrees in recent years had received better preparation to develop test items. Still another possibility is that this could be a spurious finding due to the small sample size (15 teachers) or some other unknown error in sampling.

This investigation did not examine the validity of teachers' total tests. It was limited to study of individual items. Often, when an item was judged to lack face validity, another item for an adjacent objective was better suited and the pair of items together was valid to test the two objectives. This informal finding would be difficult to quantify and demonstrate. However, since 85% of the items were judged to have good face validity and only 4% were judged to be invalid, if any sizeable portion of the remaining 10% (judged marginally

valid) were in fact usefully valid or could become valid when switched with neighboring items on the same test, then it would be safe to conclude that these technology teachers can develop reasonably valid tests.

Previous research has shown tests to be time on-task activities which promote learning of the subject matter tested. One criticism of teacher-made tests has been that they waste time. If the tests are good ones then much of the time devoted to them may be well spent. However, poorly developed tests would still be a waste of time for learning and evaluation purposes. This study identified several weaknesses in test items developed by teachers. Other factors, such as selection of different types of items for differing objectives, total test validity, problems in scoring and grading, instructions to students about tests, and others could not be addressed in this particular study--but they remain as important research problems. These questions need to be answered before meaningful conclusions can be drawn about the learning value of time students spend taking teacher-made tests.

It is concluded that technology teachers could be better prepared to develop tests if more of them were required to take a testing and measurements course. It is also concluded that the teachers in this sample are generally capable of developing valid test items, but that the items teachers prepare vary in the 9 aspects of overall quality as predicted by previous research.

References

Burdin, J.L. (1982). Teacher certification. In H.E. Mitzel (Ed.), *Encyclopedia of education research* (5th ed.). New York: Free Press.

Carter, K. (1984). Do teachers understand the principles for writing tests? *Journal of Teacher Education*, *35*(6), 57-60.

DeLuca, V.W. & Haynie, W.J. (1990). *Updating, computerization, and field validation of competency-based test-item banks for selected construction and communications technology education courses* (Contract No. RFP 90-A-07). Raleigh, NC: North Carolina State Department of Public Instruction.

DeLuca, V.W. & Haynie, W.J. (1989). *Updating, computerization, and field validation of competency-based test-item banks for selected manufacturing technology education courses* (Contract No. RFP 88-R-03). Raleigh, NC: North Carolina State Department of Public Instruction.

Fleming, M. & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway (Ed.), *Testing in the schools: New directions for testing and measurement, No. 19* (pp.29-38). San Francisco: Jossey-Bass.

Gullickson, A.R. (1982). *Survey data collected in survey of South Dakota teachers' attitudes and opinions toward testing*. Vermillion: University of South Dakota.

Gullickson, A.R. & Ellwein, M.C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. *Educational Measurement: Issues and Practice*, *4*(1), 15-18.

Haynie, W.J. (1983). *Student evaluation: The teacher's most difficult job*. Monograph Series of the Virginia Industrial Arts Teacher Education Council, Monograph Number 11.

Haynie, W.J. (1987). *Anticipation of tests as a learning variable*. Unpublished manuscript, North Carolina State University, Raleigh, NC.

Haynie, W.J. (1990). Effects of tests and anticipation of tests on learning via videotaped materials. *Journal of Industrial Teacher Education, 27*(4), 18-30.

Haynie, W.J. (1991). *Effects of take-home and in-class tests on delayed retention learning acquired via individualized, self-paced instructional texts*. Manuscript submitted for publication.

Herman, J. & Dorr-Bremme, D.W. (1982). *Assessing students: Teachers' routine practices and reasoning*. Paper presented at the annual meeting of the American Educational Research Association, New York.

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

Jackson, S.D. (1987). *The relationship between time and achievement in selected automobile mechanics classes*. (Doctoral dissertation, Texas A&M University).

Mehrens, W.A. & Lehmann, I.J. (1984). *Measurement and Evaluation in Education and Psychology*. 3rd ed. New York: Holt, Rinehart, and Winston.

Mehrens, W.A. & Lehmann, I.J. (1987). Using teacher-made measurement devices. *NASSP Bulletin, 71*(496), 36-44.

Newman, D.C. & Stallings, W.M. (1982, March). *Teacher competency in classroom testing, measurement preparation, and classroom testing practices*. Paper presented at the Annual Meeting of the National Council on measurement in Education. (In Mehrens & Lehmann, 1987)

Nungester, R.J. & Duchastel, P.C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74*(1), 18-22.

Salmon, P.B. (Ed.). (1982). *Time on task: Using instructional time more effectively*. Arlington, VA: American Association of School Administrators.

Seifert, E.H. & Beck, J.J. (1984). Relationships between task time and learning gains in secondary schools. *Journal of Educational Research, 78*(1), 5-10.

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. & Bridgeford, N.J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice, 5*(2), 5-17.