DOCUMENT RESUME

ED 141 396 TM 006 345

AUTHOR Petrosko, Joseph M.
TITLE The Quality of High School Reading and Vocabulary Tests: Implications for the Researcher.
PUB DATE [Apr 77]
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Evaluation; Evaluation Criteria; Norms; Predictive Validity; *Reading Comprehension; *Reading Tests; Secondary Education; *Standardized Tests; Test Interpretation; Test Reliability; Test Validity; *Vocabulary
IDENTIFIERS MEAN Evaluation System

ABSTRACT

Three hundred-fifty-two standardized tests of reading comprehension and 373 standardized vocabulary measures were analyzed in terms of a number of criteria related to psychometric quality and educational ability. The criteria were based primarily on the Standards for Educational and Psychological Tests developed by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education, and the MEAN test evaluation system developed by the Center for the Study of Evaluation. Fewer than 10 percent of the tests reported reliability and validity coefficients sufficiently high to make them appropriate for use by researchers or evaluators. Approximately 30 to 40 percent of the tests reported good raw score distribution characteristics and useful converted scores. Very few instruments had nationally representative norm samples or useful information for decision making about pupils. Given the quality of many "off-the-shelf" instruments, researchers and evaluators should have a variety of alternate measurement strategies at hand. (Author/MV)

The Quality of High School Reading and Vocabulary Tests: Implications for the Researcher*

Joseph M. Petrosko
University of Louisville

*Paper presented at the annual convention of the American Educational Research Association, New York City, April 1977.

ABSTRACT

The Quality of High School Reading and Vocabulary Tests: Implications for the Researcher

Joseph M. Petrosko
University of Louisville

Standardized tests in the areas of Reading Comprehension (N=352) and Vocabulary (N=373) were analyzed on a number of criteria related to educational and psychometric quality. For many criteria related to validity and reliability, fewer than 10% of the tests reported correlations sufficiently high enough to make them strong candidates for use by a researcher or evaluator. A moderately large percentage (in the area of 30 to 40 percent) of tests had good raw score distribution characteristics and useful converted scores. Very few tests had nationally representative norm samples or useful information for decision-making about pupils. The researcher needs to have a variety of measurement strategies in mind to compensate for the weaknesses of many "off-the-shelf" instruments.

In 1974, a large scale project was completed that involved a quality assessment of all published standardized tests aimed at secondary level students (Hoepfner, Conniff, Petrosko, Watkins, Erlich, Todaro, Hoyt, McGuire, Klibanoff, Stangel, Lee, Rest, Hufano, Bastone, Ogilvie, Hunter, & Johnson, 1974). Approximately 5,400 tests (or subtests of larger test batteries) were subjected to a detailed evaluation procedure. Tests were rated on many criteria of psychometric and educational quality. For the most part, the criteria well represented the concerns expressed in the Standards for Educational and Psychological Tests (Joint Committee of the American Psychological Association, American Educational Research Association and National Council on Measurement in Education, 1974).

This large body of data allowed many types of comparisons to be made regarding current tests. Petrosko and Hufano (1975) examined the quality of high school mathematics tests. Shani and Petrosko (1976) used the entire set of ratings to develop a theory to guide the evaluation of standardized tests. The present study focuses on the quality of high school tests in Reading Comprehension and Vocabulary.

The objectives of the study were: 1) to report on the general level of quality of Reading and Vocabulary tests 2) to explore the implications of these findings for test users, especially researchers.

Method

How tests were evaluated

A detailed description of evaluation procedures is contained in Hoepfner et al. (1974). The following describes, in brief, the process that was followed

Following a canvass of test catalogs and test publishers, all tests suitable or recommended for secondary students (except clinical and projective measures) were ordered. For each test, evaluators decided if the instruments would be evaluated

in whole or in parts. A subtest was evaluated if it yielded a separate score which the publisher or the organization of the test itself clearly indicated could be interpreted separately. Using this rule, a test was evaluated: 1) as a whole and for each of the subtests, or 2) only as a whole, or 3) only for the subtests.

Each test and subtest was categorized by grade level according to the claims or directions of the publisher. In the absence of such information, test evaluators estimated grade levels according to common curriculum sequences and item difficulties. Tests were assigned to one or more of three separate categories: 7-8, 9-10, or 11-12. Those tests that spanned categories (e.g. some tests were labeled "high school" and intended for grades 9 through 12) were evaluated for each grade combination and reported separately at each level.

Two raters independently assigned each test or subtest to one of 298 categories - 234 goals subsumed under 64 more general goals. Developed after consulting textbooks, curriculum guides, journal articles, and other publications, the goals constituted a comprehensive taxonomy of secondary education in terms of student outcomes. The wide ranging collection included traditional subject-matter areas (e.g. goals in English, Mathematics, and Science), Vocational and Career Education, Personality Characteristics (i.e. goals in the affective domain), and Physical Education.

After decisions were made about evaluation of subtests, about assignment to grade level, and about categorization into goal area, the tests were evaluated on 39 criteria of test quality. The 39 criteria were grouped into four broad areas: Measurement Validity, Examinee Appropriateness, Administrative Usability, and Normed Technical Excellence (yielding the acronym MEAN evaluation system). These criteria were applied only to the materials provided by the test publisher or distributor.

For each test or subscale that was evaluated, the reviewer used a standard rating form. Every test was independently rated according to the MEAN system by at least two raters, each working without access to the other's ratings. The final adjudication of test assignment to goal area and adjudication of the 39 quality ratings were both performed by an additional rater. All raters had the same information on each test--a standard specimen set consisting of the test itself and, in some cases, a technical manual or other types of supporting information.

It is important to point out that a standard was applied in considering supporting information on all tests. Thirteen of the 39 MEAN criteria deal with empirical aspects of tests, mostly related to validity and reliability. For these criteria, two rules were devised: The student samples used in generating empirical data must: (1) contain some students in at least one of the two grades for a given evaluation (7-8, 9-10, 11-12) and (2) must include students at, but not more than one-grade level above or below these grades. Using these rules, a test being evaluated for Grades 9-10 would receive credit for validity or reliability criteria if student samples contained any grade combination that included grade 9 and grade 10, but did not include any students at grade 7 or below or grade 12 and above.

The practical effect of these rules was to downgrade those tests where care was not taken in reporting data or in planning validity and reliability studies. A number of tests had "high school" forms in which a mix of students from all grade levels of high school were used in test development. Such data were not credited. For example, the data for the grades 9-10 evaluation did not receive credit because grade 12 is more than one grade above grade 10. Similarly, the data for grades 11-12 were not credited since grade 9 is more than one grade below 11.

Test Evaluation Personnel

All test evaluations were performed by individuals trained in educational testing. The majority of test evaluators possessed either an MA or a Ph.D. in education or psychology.

Goal Area Selected for Study

For this study, the tests categorized into goal areas representing reading comprehension and vocabulary skills were examined. The description of these goals are as follows.

Goal 6 A

Reading Comprehension Skills

Identifies the main idea and important details; determines the meaning of words from the way they are used; applies the reading technique appropriate to the subject matter. Draws inferences from material read.

Goal 25 A

Comprehension and Production of Information (Vocabulary)

Has a broad vocabulary. Produces needed information and abstract ideas. Describes pictures or sounds and illustrates ideas with other ideas.

As might be discerned from a careful reading of goal 25A, this is a broad area—covering some measures labeled "intelligence" tests. However, the majority of tests falling in the goal area were traditional vocabulary tests, for example the vocabulary subtests of achievement test batteries. The tests eliminated from the goal area for this analysis were the Gilliard Learning Potential Examination (Picture Completion Subtest), the scales of the Goodenough-Harris Drawing Test, the Hiskey-Nebraska Test of Learning Aptitude (Completion of Drawings Subtest) and the Mathematical and Technical Test (Completing Pictures Subtest). What remained were a large number of measures that all had as their objective a tapping of student skills in determining word meaning.

Tests covering high-school (i.e. grades 9-12) were analyzed in this study. For Reading Comprehension 352 test and subtest evaluations were analyzed; for Vocabulary, the total number was 373.

Results

Several criteria in the MEAN system dealt with the quality of a test's item selection procedures. Table 1 shows how tests fared on two separate criteria in this area. For 51% of the Reading Comprehension tests and 65% of the Vocabulary tests, no information was given by the publisher on item selection procedures. It was impossible to determine from where items were derived--textbooks, curriculum plans or some other source were not cited. For the criterion related to empirical item selection procedures--fewer than 10% of tests provided evidence that procedures like item analysis or criterion groups analysis were used.

Table I

Numbers and Percentages of Tests Rated for Quality of Item Selection

	Reading Comprehension		Vocabulary	
	N	%	N	%
Item Selection Sources				
Detailed Description of Item Selection	63	18%	24	6%
Statement Made on Item Selection	110	31%	107	29%
No Information on Item Selection	179	51%	242	65%
Empirical Procedures for Item Selection				
Evidence of Empirical Procedures	20	6%	35	9%
No Evidence of Empirical Procedures	332	94%	338	91%

The evaluation system covered several areas in construct validity. The latter term has special salience, of course, in personality testing where a developer of a new measure might justify such a test by empirically demonstrating its relationship with some hypothetical construct. Nevertheless, the term has meaning in achievement testing, insofar as such measures gain in usability by demonstrating their independence from other measures and their "purity" of content (in the factor analytic sense).

Table 2 shows that only about 1% of tests gave any information on divergent validity (low correlations with other measures) or reported evidence of using factor analysis in developing the test. Further, few tests (again, only about 1%) were reported as having been used in an experiment or an evaluation. A fairly large proportion of tests, however, (64% of the Reading Comprehension and 50% of the Vocabulary) did give a statement justifying the test's existence. All that was required was a sinple comment that showed that the developers had some specified educational, psychological or learning theory in mind when they developed the instrument.

Table 2

Numbers and Percentages of Tests Rated for Aspects of Construct Validity

		Reading Comprehension		Vocabulary	
		N	%	N	%
Divergent Validity	Yes	11	3%	4	1%
Information Given	No	341	97%	369	99%
Factorial Validity	Yes	2	1%	2	1%
Information Given	No	350	99%	371	99%
Experimental Use of	Yes	2	1%	3	1%
Test Reported	No	350	99%	370	99%
Theoretical Support	Yes	224	64%	188	50%
For Test Given	No	128	36%	185	50%

A very important consideration for any test relates to its concurrent and predictive validity. How well does a test relate to established measures or relate to future outcomes? Table 3 displays how language tests were rated on these criteria. About 15% of both Reading Comprehension and Vocabulary tests reported concurrent correlations greater than .70. The rest were below .70 for such correlations or reported no validity studies in this area. The situation was worse in rated predictive validity. Approximately 90% of the tests did not report such studies or reported data that were not acceptable. Regarding the latter point, in both concurrent and predictive validity, test evaluators judged the quality of the criterion. If the criterion--a test or a measure of success at something--was patently irrelevant or unrelated to the goal area of the evaluated test, the test was not credited.

Table 3

Number and Percentages of Tests Rated for Concurrent Validity and Predictive Validity

	Reading Comprehension		Vocabulary	
	N	%	N	%
Concurrent Validity				
Studies referred to $r \geq .70$	53	15%	55	15%
Studies referred to $.30 \leq r < .70$	27	8%	12	3%
No studies referred to	272	77%	306	82%
Predictive Validity				
$r \geq .70$, Relevant criteria, Interval of ≥ 1 month, cross-validation shrinkage $\leq 10\%$	0	0%	0	0%
$r \geq .70$, Relevant criteria, Interval of ≥ 1 month	7	2%	6	2%
$.30 \leq r < .70$ or Questionable Criteria	29	8%	22	6%
No study performed or Irrelevent Study	316	90%	345	92%

An critical element of quality in any standardized test concerns its reliability. How consistent are scores obtained by students? Table 4 shows ratings in three types of test reliability. The pattern of results was remarkably similar for both Reading Comprehension and Vocabulary. Tests were strongest in internal consistency. About 20% of rated tests had coefficients above .70. With alternate form reliability about 10% were above .70, while only 5% of test-retest correlations exceeded this benchmark figure.

For test-retest reliability, tests were credited if the time span between testing was one month or more. Retesting with the same form or delayed alternate form testing were both acceptable. Regarding the criterion of internal consistency, split-half, Kuder-Richardson, or alpha coefficients were all accepted as evidence. For alternate form reliability, either immediate or delayed testing was credited.

Table 4

Numbers and Percentages of Tests Rated for Common Types of Reliability

	Reading Comprehension		Vocabulary	
	N	%	N	%
Test-Retest Coefficient				
$r \geq .90$	2	1%	8	2%
$.80 \leq r < .90$	4	1%	8	2%
$.70 \leq r < .80$	6	2%	3	1%
$r < .70$	340	96%	354	95%
Internal Consistency Coefficient				
$r \geq .90$	37	11%	50	13%
$.80 \leq r < .90$	21	6%	25	7%
$.70 \leq r < .80$	4	1%	3	1%
$r < .70$	290	82%	295	79%
Alternate Form Coefficient				
$r \geq .90$	10	3%	19	5%
$.80 \leq r < .90$	20	6%	21	6%
$.70 \leq r < .80$	12	3%	3	1%
$r < .70$	310	88%	330	88%

A consideration was given during test evaluation procedures of various factors of a test's administrative usability. Validity and reliability are important, but do not tell the whole story. Several criteria related to test interpretation are listed in Table 5.

A larger proportion of Vocabulary tests than Reading Comprehension (19% vs. 8%) had a wide norm range. This norm range criterion was applied to determine if tests were restricted in range. The latter occurred if the upper and lower limits of the norm group were less than two years beyond the levels for which the test was evaluated. For example, a test evaluated for grades 9-10 having no 8th or 12th graders in the norm group was judged restricted in range.

Again Vocabulary tests showed superiority in score interpretation--75% had common converted scores, in contrast to 57% of Reading Comprehension tests. A surprisingly large percentage of both types of tests had novel scores, ambiguous scores, or not converted scores at all.

For the remaining three criteria in score interpretation, results were similar for tests in the two goal areas. Both types had tests with relatively straightforward procedures for conversion from raw score to converted score, both had tests with not nationally representative norm groups and both had a majority of instruments being capable of interpretation by school staff members.

Table 5

Numbers and Percentages of Tests Rated on Criteria Related to Test Interpretation

	Reading Comprehension		Vocabulary	
	N	%	N	%
Norm Range				
At least 2 years	27	8%	71	19%
Restricted range	325	92%	302	81%
Score Interpretation				
Common and simple converted scores[a]	202	57%	279	75%
Novel, ambiguous, or no converted scores	150	43%	94	25%
Score Conversion				
Simple or no conversion	247	70%	267	72%
Poor Tables or 2 step conversion	102	29%	99	27%
Complicated conversion	3	1%	7	1%
Norm Group				
Nationally representative[b]	14	4%	23	6%
Not nationally representative	338	96%	350	94%
Score Interpreter				
School Staff	346	98%	328	88%
Specialist	6	2%	45	12%

[a]Common and simple were: pass/fail, percentile ranks, mental ages, deviation IQ's, and grade equivalents.

[b]Nationally representative meant having at least four of the following attributes: (1) cluster, stratified, or random sampling; (2) norming less than five years old; (3) all areas of U.S. sampled; (4) appropriate age range represented and exhausted; (5) racial/ethnic representation or separate norms for such groups; (6) urban, suburban, and rural sampling.

To elaborate on several areas related to norm samples and quality of scores, three criteria dealt with these topics in depth. Table 6 gives percentages relevant to such concerns. It was found that 70% of Vocabulary tests, but only 57% of Reading Comprehension tests had replicability of standardization procedures. This meant that procedures of administration, scoring and interpretation were sufficiently standardized so that results could be duplicated from the norm group. About half of the tests gave no information on score distributions or reported badly skewed distributions. Thirty percent or more of tests in both areas had well drawn out score distributions.

About half of the tests considered had some type of fairly well graduated converted scale. But a dismayingly large number had crude graduation or a type of novel scale that most test users would not be familiar with.

Table 6

Numbers and Percentages of Tests Rated on Replicability of Standardization Procedures, Range of Coverage and Quality of Score Graduation

	Reading Comprehension		Vocabulary	
	N	%	N	%
Can the testing procedure be duplicated? Are procedures of administration, scoring, and interpretation standardized?				
Yes	202	57%	262	70%
No	150	43%	111	30%
Does the rest have an adequate range of coverage? (high ceiling, low floor, symmetrical distribution)				
Tails of distribution drawn out, floor or ceiling not reached	105	30%	142	38%
One tail of distribution drawn out, floor or ceiling not reached	21	6%	14	4%
Floor or ceiling reached	29	8%	23	6%
No information on score distribution or badly skewed	197	56%	194	52%
Quality of Score Graduation				
Percentiles, grade equivalents, or mental ages	132	38%	156	42%
Deciles, stanines, T-scores, or z-scores	26	7%	60	16%
Pass-fail, quartiles, or novel scales	194	55%	157	42%

The last criterion of test quality focused on how well the test helped a user make a decision about the test taker. Tests were rated high if they gave prescriptive information on a student (e.g. information associating a score with some educational placement decision). Table 7 shows that 90% of tests had, at best, poor guidelines for decisions. They had, for the most part, little information to allow the meaning of score to be translated into some action.

Table 7

Number and Percentages of Tests Rated on Decision-Making Utility

	Reading Comprehension		Vocabulary	
	N	%	N	%
Does the test provide information useful for making any individual or group decisions?				
Definite, prescriptive decisions	0	0%	1	< 1%
Suggestive decisions	30	9%	34	9%
Poor guidelines for decisions	115	33%	104	28%
Little or no information for decisions	207	59%	234	63%

Discussion

Differences between Reading Comprehension and Vocabulary

It might be well to first review the findings with respect to differences between tests in Reading Comprehension and Vocabulary. Although the pattern of percentages on the various criteria was similar for the two goal areas, in some cases substantial differences were found.

On the very first criterion considered, sources for Item Selection, a markedly larger proportion of tests in Reading Comprehension rather than Vocabulary (18 versus 6 percent) had a detailed description of Item Selection procedures. The latter meant that the publisher provided a statement on where items came from or what resources (e.g. curriculum guides) were used in arriving at an initial pool of questions.

The differences between the two goal areas may have reflected the necessity, in the case of Reading Comprehension, to clearly describe the type of source material and alert the test purchaser to the specific content areas from which reading passages would come. Such a justification was perceived as, perhaps less necessary in Vocabulary. In selecting vocabulary items, some publishers may have simply used item tryout information to eliminate extremely difficult and easy words from some arbitrary starting list.

In another point of contrast, 64% of the Reading tests and only 50% of the Vocabulary tests reported theoretical support (see table 2). This meant that more of the Reading tests gave some sort of statement of rationale--some defense for the tests existence. The reasons may well have been the same

as the relative superiority of Reading tests on the first criterion. It may have been simply that authors of Vocabulary Tests felt less need for any justification--under the assumption that the utility of knowing something about a student's knowledge of Vocabulary is obvious.

Two criteria on which Vocabulary tests made a better showing than Reading were Norm Range and Score Interpretation (see table 5). Nineteen percent of Vocabulary as against eight percent of Reading tests had a norm range of at least 2 years. The latter meant that norm groups had upper and lower limits at least 2 years beyond the levels for which the test was evaluated (i.e. levels 9-10 or 11-12). Moreover, a fairly large discrepancy existed between the two goal areas on the dimension Score Interpretation. For Vocabulary, fully 75% versus 57% of Reading Tests had common or simple converted scores (e.g. percentiles). In other words, a surprisingly large proportion of 43% of the Reading tests had novel, ambigious or no converted scores.

Vocabulary tests were relatively less superior on the criterion Score Interpreter. Twelve percent of them required a specialist to interpret, but only 2% of the Reading tests required a specially trained score interpreter.

These findings, the relative superiority of Vocabulary tests in Norm Range and Score Interpretation and their relative inferiority on the Score Interpreter criterion, may be related to the varied uses of the two types of tests. At least some of the Vocabulary tests came from batteries where they played the role of an "aptitude" measure. Given the association of vocabulary knowledge with IQ (in terms of often reported overlapping variance) it may very well be that Vocabulary tests shared some aspects with IQ measures that Reading tests did not. This might explain the superiority of tests in Vocabulary in Norm Range (the use of a wide age span for norms) and Score

Interpretation (the use of common converted scores) and the greater likelihood of a Vocabulary test requiring a specialist score interpreter.

The last criterion in which Reading Comprehension and Vocabulary differed was in the area Replicability of Standardization Procedures. Vocabulary had an advantage here. Seventy percent of its tests had replicable procedures--only 57% of Reading tests were so judged. This area related to whether the test provided uniformity of procedures for administering and scoring and whether the test user could use the test with samples similar to the standardization group. In other words,were the circumstances in standardizing the test similar to those faced by a test user in testing a typical group of students?

The general results and their implications

The results, despite some points of contrast, were fairly consistent for the two educational areas across the various criteria. These results have some implications for researchers and other test users.

1. the finding that few tests gave information on item selection reinforces the importance of the researcher carefully looking over the items themselves. One cannot expect much guidance from publishers on sources for items and, furthermore, any general statements about such sources may not be useful for many users. A content analysis is de rigueur. Unfortunately, 85% of Reading and Vocabulary tests reported low correlations for concurrent validity or reported nothing at all. The relationship between many little known Reading and Vocabulary tests and established measures is unclear.

2. A few tests had for a well specified age range, very high reliabilities (i.e. above .90), but the great majority did not. The researcher is fortunate if a test with high reliability might have enough other requisite characteristics that it can be used in a given research circumstance. If a test with less-than-optimal reliability must be used the following considerations might be kept in mind.

a) the researcher may often have to estimate reliabilities for a given (narrow) age range. Too often, publishers perform reliability studies with samples having wide age ranges.

b) Some thought might be given to performing small scale reliability studies, especially for special student populations.

c) Many tests had reliabilities below .70. Researchers using tests for evaluation purposes should be sensitive to problems of internal validity bias due to instrumentation error (Campbell and Stanley, 1964). A test with .70 reliability is one in which only about 50% of the variance is shared for the two scores (i.e. in alternate form and test-retest situations). Serious thought should be given to some measurement strategy that optimizes inferences about a program or treatment under study. Using more than one measure--the method of "converging operations" (Webb, Campbell, Schwartz, Sechrest, 1966)--is one such strategy.

3. Few tests gave clues as to what decisions could be made about individuals or groups based on test scores. This points up the necessity for thinking out in advance exactly how scores will be used. Reading and Vocabulary tests are useful for program evaluation purposes. For example, standardized secondary level tests in reading and other skills are being used to evaluate success of the Emergency School Aid Act (ESAA) program. But their utility for other purposes might at time be questioned. This is true especially given the fact that few predictive validity studies were identified for the tests examined in this study. There was little data on how tests related to such things as job performance or grade point average for the first year of college.

A Concluding Note

There were several limitations to this study. One concerns the procedure of acquiring and categorizing tests. Virtually every test on the market was obtained and evaluated. This meant that some rather obscure instruments were given ratings along with very well known tests. There is some justification for this, however. Tests are on the market because enough people buy them to allow a profit for the publisher. In the absence of information on how many tests of a given type are sold, it is a fair

assumption that hundreds (more realistically, thousands) of copies are sold every year of even little known instruments. It is a defensible proposition that these tests should be evaluated.

Another point to be made on this study's test evaluations is an issue related to the evaluation criteria themselves. The criteria were general and were applied to tests in every subject domain (the complete work by Hoepfner et al., 1974, lists 298 goal areas into which tests were categorized). The test purchaser and researcher should be aware of special criteria aimed at Reading Comprehension and Vocabulary tests exclusively. Such criteria were not included in the present report or the source data from which it was derived. But researchers should be cognizant of special problems with language oriented tests. Probably the most significant of these is the passage dependence of Reading Comprehension tests. Tuinman (1973-1974) found that some items in Reading Comprehension tests are not dependent on the passage of prose that they follow. Such items are answered correctly at a higher than chance rate by subjects who do not read the passage with which the items are ostensibly linked. Needless to say, this weakness in measurement needs to be noted by a prospective test user.

References

Campbell, D.T., & Stanley, J.C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1964.

Hoepfner, R., Conniff, W., Jr., Petrosko, J.M., Watkins, J., Erlich O., Todaro, R.S., Hoyt, M.F., McGuire, T.C., Klibanoff. L.S., Stangel, G.F., Lee, H.B., Rest, S., Hufano, L., Bastone, M., Ogilvie. V.N., Hunter. R., & Johnson, B.L. CSE secondary school test evaluations (3 vols.). Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Joint Committee of the American Psychological Association, American Educational Research Association & National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Petrosko, J.M. & Hufano, L. An assessment of the quality of high school mathematics tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., April 1975.

Shani, B. & Petrosko. J.M. Structural components derived from evaluating standardized tests. Journal of Educational Measurement, 1976, 13, 283-296.

Tuinman, J.J. Determining the passage dependency of comprehension questions in 5 major tests. Reading Research Quarterly, 1973-1974, 9, 206-223.

Webb, E., Campbell, D., Schwartz, R., & Sechrest, L. Unobtrusive measures: non-reactive research in the social sciences. Chicago: Rand McNally, 1966.