DOCUMENT RESUME

ED 109 188 TM 004 638

AUTHOR Petrosko, Joseph M.; Hufano, Linda
TITLE An Assessment of the Quality of High School Mathematics Tests.
PUB DATE [Apr 75]
NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D. C., March 31-April 2, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Algebra; Comparative Analysis; *Evaluation; *Evaluation Criteria; Geometry; *Mathematics; *Secondary Education; Senior High Schools; *Standardized Tests; Test Construction; Test Reliability; Tests; Test Validity

ABSTRACT

An assessment was made of the psychometric and educational quality of all high school level tests of general mathematics, applied mathematics, algebra and geometry. The study was part of a large-scale project involving evaluations of all standardized secondary school tests available in the United States. Assessments revealed most tests to be low in many types of validity and reliability. Tests of general mathematics, which included arithmetic, fared the best across 39 criteria of test quality. Test developers are not meeting many basic standards of test quality in constructing mathematics tests. (Author)

AN ASSESSMENT OF THE QUALITY OF HIGH SCHOOL MATHEMATICS TESTS*

Joseph M. Petrosko
Center for the Study of Evaluation
UCLA Graduate School of Education

and

Linda Hufano
Garvey School District
Rosemead, California

*Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., April, 1975

Between 1972 and 1974, the Center for the Study of Evaluation (CSE) systematically evaluated the great majority of published tests on the secondary-education level (Grades 7 through 12). Evaluations of more than 5400 tests or subtests of batteries were published in a set of three volumes (Hoepfner et al., 1974). The evaluations provide the user with 39 educational and psychometric quality ratings of secondary-level standardized tests.

This study concerns a subset of the evaluation ratings - that of mathematics tests in grades 9 through 12. The objective of the study was twofold. 1) to compare and contrast the quality of tests in various areas of mathematics, and 2) to note those aspects of test construction to which developers could direct their future efforts.

METHOD

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.

Procedure

A multi-step procedure was followed in the evaluations:

1. Following a canvass of test catalogs and test publishers, all tests suitable or recommended for secondary students, except clinical and projective measures, were ordered.

2. 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.

3. 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.

4. Two raters independently assigned each test or subtest to one of 298 goal categories - 234 goals subsumed under 64 more general goals. The goals comprised a set especially constructed by the Evaluation Technologies Program run by CSE. Using 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.

5. 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 only applied 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.

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 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 criterion 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.

The complete set of evaluation ratings, along with the list of goals and a detailed description of the evaluation procedure are contained in Hoepfner et al. (1974). The present study focuses on tests given in mathematics for students in grades 9 through 12 (i.e., contained in the volumes for grades 9-10 and 11-12). These tests were crosstabulated with a number of the 39 evaluation criteria.

Four areas of mathematics were selected for study. Their descriptions follow.

General Mathematics

Including - Arithmetic, Number Concepts Systems and Sets; Measurement.

Applied Mathematics

Including - Business and Consumer Math; Industrial and Vocational Math; Computer Programming; Computer Theory and Practice.

Algebra

Including - Algebraic Skills and Concepts, Real and Complex Number Systems; Equations and Inequalities; Exponents, Radicals, Logs, and Functions; Linear Algebra.

Geometry

Including - Informal Geometry; The Nature of Proof in Math; Euclidean Plane Geometry; Coordinate Plane Geometry; Solid Geometry.

RESULTS

The ratings of tests on several criteria related to content and construct validity are shown in Table 1. Two important aspects of content validity were examined whether item selection procedures were rigorous and whether empirical item selection occurred. For approximately 50% of the tests across all four mathematics categories no information was offered on how items were selected (evidence was sought on the publisher's sources of information for test construction - curriculum guides, textbooks etc.). Across the categories, about 10 percent or fewer of the tests contained a report of any empirical procedures for item selection (e.g. jury of experts, item analysis, criterion group analysis, etc.). As with all validity and reliability criteria, it must be remembered that empirical procedures had to be based on samples of students including, but not more than one year above or below, the age range for which the test was evaluated.

In construct validity, tests were examined on four criteria. Few reported divergent validity information (correlations), factorial validity information (factor loadings), or experimental uses of a test (employing it in experiments or evaluations). A fairly large proportion of tests in General Math and Applied Math were credited with Theoretical Support. In order to be so rated, it was required that some justification be given of the test's existence. An example of such justification might be a statement like: "in the past decade greater attention has been directed by educators to the teaching and learning of set theory as a basis for the understanding of mathematics."

Table 1

Percentages of Tests Receiving Ratings in Content and Construct Validity

		General Mathematics (N=322)	Applied Mathematics (N=26)	Algebra (N=122)	Geometry (N=52)
Content Validity					
Item Selection					
Detailed Description of Item Selection		18	7	16	14
Statement Made on Item Selection		35	8	47	27
No Information on Item Selection		47	85	37	59
Empirical Item Selection	YES	10	0	8	6
	NO	90	100	92	96
Construct Validity					
Divergent Validity Information	YES	2	0	1	0
	NO	98	140	99	100
Factorial Validity Information	YES	2	0	3	0
	NO	98	100	97	100
Experimental Use of Test	YES	1	0	2	0
	NO	99	100	98	100
Theoretical support given	YES	70	65	8	8
	NO	30	35	92	92

Table 2 shows ratings in concurrent and predictive validity. For both types of validity and across the four areas of mathematics, few studies of any kind were reported, although a fair number of General

Math tests had concurrent validity correlations above .70. Tests in Applied Mathematics were little better in predictive validity than the other math areas, even though the Applied Math area was more clearly related to immediate post-high-school employment. Presumably the latter fact would make collection of data on some criterion such as job success a relatively straightforward process.

For both concurrent and predictive validity, referred to by the Standards for Educational and Psychological Tests (1974) as *criterion-related validities*, test evaluators judged the quality of the criterion itself. 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 2

Percentages of Tests Receiving Ratings in Concurrent and Predictive Validity

	General Mathematics (N-322)	Applied Mathematics (N=26)	Algebra (N=122)	Geometry (N=52)
Concurrent Validity				
Studies referred to $r \geq .70$	15	0	6	4
Studies referred to $.30 \leq r < .70$	2	0	3	2
No studies referred to	83	100	91	94
Predictive Validity				
$r \geq .70$, Relevant criteria, Interval of ≥ 1 month, cross-validation shrinkage $\leq 10\%$	1	0	0	2
$r \geq .70$, Relevant criteria, Interval of ≥ 1 month	3	4	3	2
$.30 \leq r < .70$ or Questionable Criteria	5	8	3	2
No study performed or Irrelevant Study	92	88	94	94

Table 3 shows how tests fared in reported correlations of test-retest, internal consistency, and alternate-form reliability. For well over 75% of the tests, correlations fell below .70 or were not reported. A fair percentage (19%) of General Mathematics tests had high internal consistency coefficients.

For test-retest reliability, tests were credited if the time span between testings 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 3

Percentages of Tests Receiving Ratings in Three Types of Reliability

	General Mathematics (N=322)	Applied Mathematics (N=26)	Algebra (N=122)	Geometry (N=52)
Test-Retest Coefficient				
$r \geq .90$	1	0	0	0
$.84 \leq r < .90$	4	0	0	0
$.70 \leq r < .80$	2	0	0	0
$r < .70$	93	100	100	100
Internal Consistency Coefficient				
$r \geq .90$	19	0	5	0
$.80 \leq r < .90$	8	4	10	11
$.70 \leq r < .80$	0	11	1	0
$r < .70$	73	85	84	89
Alternate Form Coefficient				
$r \geq .90$	3	4	0	0
$.80 \leq r < .90$	5	8	1	0
$.70 \leq r < .80$	2	0	1	0
$r < .70$	90	89	98	100

Up to this point, the ratings were related to purely technical qualities of the test. However, many of the criteria in the MEAN test evaluation system pertain to broader issues such as (1) test interpretation, (2) quality of score distribution, and (3) utility of a test for decision making.

1. Table 4 contains five criteria related to test interpretation. First on the positive side most tests showed their capability of being interpreted by the school staff rather than by a specialist. Further, score convers on was usually simple (one step from raw score to scaled score) and 50 percent or more of the tests had commonly used converted scores, such as percentile ranks or grade equivalents. Less positive were the findings shown in Norm Range. Most tests were restricted in range, that is, 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, most tests evaluated for grades 9-10 had norm groups that did not contain 8th graders or 12th graders. Also, norm groups were rarely nationally representative and failed to achieve geographical representation or to use random sampling procedures. (See Table 4, page 10.)

2. Several other criteria on norming procedures and scores are worthy of a separate Table. As can be seen in Table 5, about two-thirds of mathematics tests had replicable testing procedures. In other words, procedures of administration, scoring, and interpretation were sufficiently standardized so that results could be duplicated or replicated from the norm group. Quality of score distribution and of score graduation varied among the areas of mathematics. About 75 percent of the Algebra and Geometry tests had badly skewed distributions (or no information available at all) and had rather crude converted scores such as quartiles. Tests in General Math and Applied Math tended to have better score distributions and more graduated standard scores. (See Table 5, page 11.)

Table 4

Percentages of Tests Receiving Ratings on Criteria Related to Test Interpretation

	General Mathematics (N=322)	Applied Mathematics (N=26)	Algebra (N=122)	Geometry (N=52)
Norm Range				
At least 2 years	15	4	1	0
Restricted range	85	96	99	100
Score Interpretation				
Common and simple converted scores[a]	62	81	51	52
Novel, ambiguous, or no converted scores	38	19	49	48
Score Conversion				
Simple or no conversion	77	81	82	83
Poor Tables or 2 step conversion	20	19	17	17
Complicated conversion	2	0	1	0
Norm Group				
Nationally representative[b]	8	0	2	2
Not nationally representative	92	100	98	98
Score Interpreter				
School staff	98	96	100	100
Specialist	2	4	0	0

[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.

Table 5

Percentage of Tests Receiving Ratings on Criteria of Replicability of Standardization Procedures, Range of Coverage, and Quality of Score Graduation

	General Mathematics (N=322)	Applied Mathematics (N=26)	Algebra (N=122)	Geometry (N=52)
Can the testing procedure be duplicated? Are procedures of administration, scoring, and interpretation standardized?				
YES	67	73	76	58
NO	33	27	24	42
Does the test have an adequate range of coverage? (High ceiling, low floor, symmetrical distribution)				
Tails of distribution drawn out, floor or ceiling not reached	27	23	15	13
One tail of distribution drawn out, floor or ceiling not reached	4	15	3	2
Floor or ceiling reached	17	12	5	2
No information on score distribution or badly skewed	52	50	78	83
Quality of Score Graduation				
Percentiles, grade equivalents, or mental ages	40	42	22	23
Deciles, stanines, T-scores, or Z-scores	16	8	6	2
Pass-fail, quartiles, or novel scales	44	50	72	75

3. A final criterion, one well worth looking at since it impinges on the reality of the school world in such a direct way, is the decision-making utility of a test. How well does a test "map" the range of scores into the domain of decisions about the educational fate of a student? Table 6 shows that few tests give prescriptive decision-making information (e.g., "a score of 30 or more means that the student will very likely succeed if channeled into introductory algebra"). Few tests in Applied Math, an area presumably involving skills useful in post-high school vocations, yielded *any* information for decisions.

Table 6

Percentage of Tests Receiving Ratings on Criterion of Decision-Making Utility

	General Mathematics (N=322)	Applied Mathematics (N=26)	Algebra (N=122)	Geometry (N=52)
Does the test provide information useful for making any individual or group decisions?				
Definite, prescriptive decisions	2	0	0	6
Suggestive decisions	27	0	19	23
Poor guidelines for decisions	20	8	35	42
Little or no information for decisions	51	92	46	29

DISCUSSION

Where mathematics tests fail

This survey of high-school mathematics tests revealed many tests to be deficient in basic aspects of test quality. The deficiencies extended across four major curriculum areas and many criteria for judging a test. A prime example concerns the criterion of content validity - a *sine qua non* of achievement tests. The present study revealed that fewer than 20 percent of secondary-math tests gave a detailed description of item-selection procedures. Remarkably few made even a general statement on item selection (e.g., "current textbooks were surveyed").

This is far from the requirement expressed in Standards for Educational and Psychological Tests (1974), where it is deemed *essential* that such information be provided: "If test performance is to be interpreted as a representative sample of performance in a universe of situations, the test manual should give a clear definition of the universe represented and describe the procedures followed in the sampling from it." (p. 45)

It is not altogether cynical to consider the poor results in light of the unchanging economics of test development. There is no way to escape the realization that ratings tended to be higher for those criteria where it was relatively cheap and easy to provide the information. This "real" fact holds true across all types of tests. Since most types of validity studies require the expense of administering other tests or the collecting of data on some criterion, the work was simply not done. For reliability (note Table 2), ratings were best for internal consistency reliability - a coefficient that requires only one test administration. The inference is inescapable.

Comparisons of tests across the four areas of mathematics may seem complicated by the widely divergent numbers in the categories. There were more General Mathematics tests than the other three areas combined. However, it is important to consider that virtual populations of tests are being examined not samples from populations. Any comparisons of General Mathematics, Applied Mathematics, Algebra, and Geometry can be assumed to pertain to the entire populations of these types of standardized secondary tests. In that sense, all percentage differences among the groups are "significant differences", although not necessarily *practically* significant. Practical significance depends on the value assumptions of the reader.

Applying the arbitrary standard of 10 percentage points being practically significant one can make a few general statements about tests in the various areas. In comparison with Applied Mathematics, Algebra, or Geometry, a larger percentage of General Mathematics tests had: high concurrent validity, high internal consistency, and a norm range covering at least two years.

Seventy percent of the General Mathematics and only eight percent of the Algebra and Geometry tests were rated as having "Theoretical Support," but this result must be interpreted carefully. The MEAN criterion of Theoretical Support had the most saliency for tests in the affective area, where a theoretical construct was inherent in the goal statement (e.g. self concept, emotional security). With achievement tests the criterion reflected a concern that some kind of statement justified the test's existence (and not necessarily with evidence supporting the statement). Many General Mathematics tests were, in effect, arithmetic tests. There are

many such instruments on the market and most teachers can easily write arithmetic items, so publishers may have felt more necessity for providing a rationale for such tests.

Tests of Algebra and Geometry tended to have a greater number of tests with poor range of coverage (inadequate floor and ceiling etc.) and with crudely graduated scores, such as quartiles. Both phenomena were undoubtedly caused by factors such as small sample sizes in norm groups and lack of rigor in term-revision procedures.

How tests can be improved

The reader who expects a startlingly innovative declaration of how Math Tests can be improved will be disappointed. If test developers carefully applied the existing technology of test construction, there would be a great improvement in instrumentation. If one had to prescribe where the efforts of test developers could be directed, the general answer would be to conduct more validity and reliability studies.

The perennially obvious requirement of content validity does not seem to be taken seriously by many publishers. Too few developers carefully define the skills they are purporting to measure and then sample items from a universe of such skills. Furthermore, a greater number of tests should have nationally representative norm samples, better score distributions, and more discriminative types of standard scores. Quartile scales based on all-white suburban samples simply do not do the job for many test purchasers. And finally, tests should relate to the real world. For example, every test in Applied Mathematics should have some type of predictive validity information preferably in terms of job performance.

The state of testing in an educational area does not exist in a vacuum. It both affects and is affected by the state of the curriculum. So not only would a clearer conception of mathematics lead to better tests, but better tests may well lead to clearer conceptions of mathematics.

Mathematics, like many other parts of the school curriculum, began undergoing close examination about 15 years ago. New curricula were developed unfortunately not always with the firmest empirical basis. Much of the problem lay with inability to measure the various skill areas in mathematics. For example, Romberg (1969) noted: "It is safe to generalize that in most mathematics studies conducted during the 1960's, researchers used inappropriate or inadequate measuring devices to assess mathematics achievement." (p. 482)

When new curricula have been compared with traditional approaches, the efforts have been hobbled by weaknesses in tests and testing programs. This point is well brought out by Walker and Schaffarzick (1974) in a review of research studies where old and new curricula were compared: "The most important shortcoming of conventional achievement tests and the most serious single limitation of comparative curricular studies done so far is the restricted range of outcomes measured." (p.106) Conventional tests tend to measure conventional outcomes, and then without the degree of validity and reliability that would form the best evidence for decision making about those being tested.

The pioneer efforts of the National Longitudinal Study of Mathematical Abilities (NLSMA) are cited by Walker and Schaffarzick (1974), Dessart and Frandsen (1973) and others as a positive example of what can be done in curriculum and test construction. Test items were carefully linked with the content areas of mathematics which in turn were linked with four main elements of achievement: Computation, Comprehension Application, and Analysis. The state of mathematics testing can only be improved if other researchers will make similar efforts. We need tests that are relevant to the needs of educators and possess the technical quality necessary for sound research.

References

American Psychological Association. Standards for Educational and Psychological Tests. Washington, D.C.: APA, 1974.

Dessart, D.J., & Frandsen, H. Research on teaching secondary school mathematics. In R.M.W. Travers (Ed.) Second Handbook of Research on Teaching. Chicago: Rand McNally, 1973.

Hoepfner, R., Conniff Jr., W.A., et al. CSE Secondary School Test Evaluations. Los Angeles: Center for the Study of Evaluation, University of California, 1974. 3 vols.

Romberg, T.A. Current research in mathematics education. Review of Educational Research, 1969, 39, 473-491.

Walker, D.F., & Schaffarzick, J. Comparing curricula. Review of Educational Research, 1974, 44, 83-111.