DOCUMENT RESUME

ED 058 314 TM 001 029

AUTHOR	Hoepfner, Ralph
TITLE	Characteristics of Standardized Tests as Evaluation Instruments.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Cooperative Research Program.
PUB DATE	Sep 71
NOTE	6p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Educational Needs; Educational Objectives; *Elementary Schools; *Evaluation Criteria; *Evaluation Techniques; Factor Analysis; Grade 1; Grade 3; Grade 5; Grade 6; Norm Referenced Tests; *Standardized Tests; Testing Problems; Test Reliability; Test Validity

ABSTRACT

All measures presently available for first, third, fifth, and sixth grade testing were evaluated according to the MEAN procedure which reflects measurement validity, examinee appropriateness, administrative usability, and normed technical excellence. Major shortcomings of these measures are presented. A factor analysis revealed four dimensions upon which the tests actually vary: usability, norm quality, focus, and psychometric quality. (MS)

UCLA CSE

Evaluation comment

Center for the Study of Evaluation

September, 1971 Vol. 3, No. 1

STATEMENT OF INTENT

The Center for the Study of Evaluation was founded in June, 1966. It is an educational research and development center sponsored by the U.S. Office of Education under the Cooperative Research Act and is the only federally funded center working exclusively on problems in educational evaluation.

The mission of the Center is to produce new evaluation materials, practices, and knowledge which can be adopted and implemented by educational agencies. Emphasis is placed on developing procedures and methodologies needed in the practical conduct of evaluation studies and on developing generalizable concepts and approaches to evaluation problems that are relevant to different levels of education. The Center is directed by Marvin C. Alkin and is staffed by an interdisciplinary team which includes specialists in education, measurement, sociology, economics, and administration.

Evaluation Comment provides discussion of significant ideas and controversial issues in the study of evaluation of educational systems and programs. A copy of *Evaluation Comment* is distributed free of charge to each scholar, researcher, or practitioner on our mailing list. One to five copies may be obtained free of charge; however, where greater quantities are needed readers are encouraged to reproduce the *Comment* themselves. To be placed on our mailing list or to order, subject to availability, additional copies of *Evaluation Comment*, please write to:

James Burry, Editor
Evaluation Comment
Center for the Study of Evaluation
145 Moore Hall
University of California, Los Angeles
Los Angeles, California 90024

IN THIS ISSUE

CHARACTERISTICS OF STANDARDIZED TESTS AS EVALUATION INSTRUMENTS

Ralph Hoepfner

University of California, Los Angeles

For years various professional organizations in education and psychology have recognized the need to set specific criteria for assessment devices. However, attempts to develop such criteria have been, at best, timid (viz.: *Technical Recommendations*). This timidity where "angels dare not tread" may not be completely reprehensible; it is the result of several factors:

(a) any set of criteria will not be equally appropriate for all types of measures,

(b) the direct result of the development of such a set of criteria would be the ability to evaluate critically all available assessment devices,

(c) the producers of the instruments might not be too pleased and, worse, might take well-reasoned issue with the criteria and their developers, and

(d) the authors, being motivated primarily by altruism and social justice, might have to take their own inadequate, but lucrative, products off the market.

The Center for the Study of Evaluation, in order to provide an equable appraisal of the output measures published for use in evaluating elementary schools, programs, and students, developed (1) a comprehensive objectives-based classification of needs-assessment areas for elementary education, and (2) a critical test evaluation procedure to apply to measurement devices in any of the need areas. Preparatory to the evaluations, all those measures presently available for elementary school evaluation at the first, third, fifth, and sixth grades were located. Each test or sub-scale was assigned to the pre-established goal area into which it best fit.

Table 1

OUTLINE OF 145 GOALS OF ELEMENTARY SCHOOL EDUCATION

AFFECTIVE
1. TEMPERAMENT: PERSONAL
   - A. Shyness-Boldness
   - B. Neuroticism-Adjustment
   - C. General Activity-Lethargy
2. TEMPERAMENT: SOCIAL
   - A. Dependence-Independence
   - B. Hostility-Friendliness
   - C. Socialization-Rebelliousness
3. ATTITUDES
   - A. School Orientation
   - B. Self Esteem
4. NEEDS AND INTERESTS
   - A. Need Achievement
   - B. Interest Areas

ARTS-CRAFTS
5. VALUING ARTS AND CRAFTS
   - A. Appreciation of Arts and Crafts
   - B. Involvement in Arts and Crafts
6. PRODUCING ARTS AND CRAFTS
   - A. Representational Skill in Arts and Crafts
   - B. Expressive Skill in Arts and Crafts
7. UNDERSTANDING ARTS AND CRAFTS
   - A. Arts and Crafts Comprehension
   - B. Developmental Understanding of Arts and Crafts

COGNITIVE
8. REASONING
   - A. Classificatory Reasoning
   - B. Relational-Implicational Reasoning
   - C. Systematic Reasoning
   - D. Spatial Reasoning
9. CREATIVITY
   - A. Creative Flexibility
   - B. Creative Fluency
10. MEMORY
   - A. Span and Serial Memory
   - B. Meaningful Memory
   - C. Spatial Memory

FOREIGN LANGUAGE
11. FOREIGN LANGUAGE SKILLS
   - A. Reading Comprehension of a Foreign Language
   - B. Oral Comprehension of a Foreign Language
   - C. Speaking Fluency in a Foreign Language
   - D. Writing Fluency in a Foreign Language
12. FOREIGN LANGUAGE ASSIMILATION
   - A. Cultural Insight through a Foreign Language
   - B. Interest in and Application of a Foreign Language

LANGUAGE ARTS
13. LANGUAGE CONSTRUCTION
   - A. Spelling
   - B. Punctuation
   - C. Capitalization
   - D. Grammar and Usage
   - E. Penmanship
   - F. Written Expression
   - G. Independent Application of Writing Skills
14. REFERENCE SKILLS
   - A. Use of Data Sources as Reference Skills
   - B. Summarizing Information for Reference

MATHEMATICS
15. ARITHMETIC CONCEPTS
   - A. Comprehension of Numbers and Sets in Mathematics
   - B. Comprehension of Positional Notation in Mathematics
   - C. Comprehension of Equations and Inequalities
   - D. Comprehension of Number Principles
16. ARITHMETIC OPERATIONS
   - A. Operations with Integers
   - B. Operations with Fractions
   - C. Operations with Decimals and Percents
17. MATHEMATICAL APPLICATIONS
   - A. Mathematic Problem Solving
   - B. Independent Application of Mathematical Skills
18. GEOMETRY
   - A. Geometric Facility
   - B. Geometric Vocabulary
19. MEASUREMENT
   - A. Measurement Reading and Making
   - B. Statistics

MUSIC
20. MUSIC APPRECIATION AND INTEREST
   - A. Music Appreciation
   - B. Music Interest and Enjoyment
21. MUSIC PERFORMANCE
   - A. Singing
   - B. Musical Instrument Playing
   - C. Dance (Rhythmic Response)
22. MUSIC UNDERSTANDING
   - A. Aural Identification of Music
   - B. Music Knowledge

PHYSICAL EDUCATION — HEALTH — SAFETY
23. HEALTH AND SAFETY
   - A. Practicing Health and Safety Principles
   - B. Understanding Health and Safety Principles
   - C. Sex Education
24. PHYSICAL SKILLS
   - A. Muscle Control (Physical Education)
   - B. Physical Development and Well-Being (Physical Education)
25. SPORTSMANSHIP
   - A. Group Activity — Sportsmanship
   - B. Interest in and Independent Participation in Sports and Games
26. PHYSICAL EDUCATION
   - A. Understanding of Rules and Strategies of Sports and Games
   - B. Knowledge of Physical Education Apparatus and Equipment

READING
27. ORAL-AURAL SKILLS
   - A. Listening Reaction and Response
   - B. Speaking
28. WORD RECOGNITION
   - A. Phonetic Recognition
   - B. Structural Recognition
29. READING MECHANICS
   - A. Oral Reading
   - B. Silent Reading Efficiency
30. READING COMPREHENSION
   - A. Recognition of Word Meanings
   - B. Understanding of Ideational Complexes
   - C. Remembering Information Read
31. READING INTERPRETATION
   - A. Inference Making from Reading Selections
   - B. Recognition of Literary Devices
   - C. Critical Reading
32. READING APPRECIATION AND RESPONSE
   - A. Attitude toward Reading
   - B. Attitude and Behavior Modification from Reading
   - C. Familiarity with Standard Children's Literature

RELIGION
33. RELIGIOUS KNOWLEDGE
34. RELIGIOUS BELIEF

SCIENCE
35. SCIENTIFIC PROCESSES
   - A. Observation and Description in Science
   - B. Use of Numbers and Measures in Science
   - C. Classification and Generalization in Science
   - D. Hypothesis Formation in Science
   - E. Operational Definitions in Science
   - F. Experimentation in Science
   - G. Formulation of Generalized Conclusions in Science
36. SCIENTIFIC KNOWLEDGE
   - A. Knowledge of Scientific Facts and Terminology
   - B. The Nature and Purpose of Science
37. SCIENTIFIC APPROACH
   - A. Science Interest and Appreciation
   - B. Application of Scientific Methods to Everyday Life

SOCIAL STUDIES
38. HISTORY AND CIVICS
   - A. Knowledge of History
   - B. Knowledge of Governments
39. GEOGRAPHY
   - A. Knowledge of Physical Geography
   - B. Knowledge of Socio-Economic Geography
40. SOCIOLOGY
   - A. Cultural Knowledge
   - B. Social Organization Knowledge
41. APPLICATION OF SOCIAL STUDIES
   - A. Research Skills in Social Studies
   - B. Citizenship
   - C. Interest in Social Studies

An outline of the goals is provided in Table 1 above. The tests and subtests were then evaluated in order to identify and endorse those output measures most appropriate, effective, and useful in assessing schools or students. The evaluation form used throughout the test evaluations is shown in Figure 1.

The MEAN (an acronym for the four criterion areas to follow) evaluation procedure critically reflects four vital areas of concern to test users: Measurement Validity, Examinee Appropriateness, Administrative Usability, and Normed Technical Excellence. Twenty-four separate evaluations, comprising the four major criterion areas, were performed on 1,649 scales independently by at least two evaluators. These scales comprise all the output measures that are prepared for or are potentially useful for evaluations within the elementary school and that are generally available to educators and researchers.

The four criteria comprising the MEAN system are explained below. They were meant to address the interest areas of educators and also of educational researchers. However, the final ratings obtained for each test indicate its appropriateness for school evaluation settings rather than for clinical or research problems.

Measurement Validity. Evaluations on the criterion of measurement validity were made in answer to the question: "Does the test appear to measure the specific educational objective?" *(entry 1 of Table 2).* This is essentially a question of content and face validity, the validities being keyed to the pre-established goal areas for elementary education. Trained evaluators were instructed to judge each test according to its capacity to assess the particular goal which it purported to measure or which a plurality of its items appeared to reflect. The judgments were made on the basis of careful reading of the items to determine whether they appeared to assess the goal and whether they proportionately assessed the whole range of content within the goal. Such judgments were fairly well structured and reliable in the content achievement areas, but were more difficult to make in the non-content areas of affective and cognitive behaviours. A second aspect of measurement validity concerned the extent of reported empirical validation, either predictive or concurrent *(entry 2, Table 2).*

Examinee Appropriateness. The second criterion of the MEAN evaluations was designed to assess how appropriate the test is for the students who will be assessed by it. Concern was directed toward the appropriateness of the test's level of comprehension, its physical format, and its required response mode.

Evaluation of the appropriateness of test content centered upon the difficulty of the semantic or numerical items and also upon the relevance or interest-arousing aspects of the items *(entry 3, Table 2).* Similar criteria were applied to the test instructions since they determine whether or not the examinee will be able to manifest his mastery of the item content *(entry 4, Table 2).* Instructions which appear simple to adults were often found to be confusing to young children. The second major area where appropriateness is felt to be impor-

MEAN TEST EVALUATION FORM

Test Name _____ Form _____ Rater _____ Date _____

Evaluation Criteria	Rating (circle one number in each row)						
1. Measurement Validities a. Content and Construct	0 (only in name)	2 (a few)	4 (some)	6 (fair job)	8 (best available)	10 (hit nail on the head)	M Total
b Concurrent and Predictive	0 (none reported)	1 (very little)	2 (some)	3 (not enough)	4 (considerable)	5 (exhaustive)	Grade
2. Examinee Appropriateness a. Comprehension: content	inappropriate 0	doubtful 1	possibly appropriate 2	probably appropriate 3	exactly right 4		
instructions	0	1	2	3	4		
b. Format 1. Visual principles	0 (complicated)		1 (probably good)		2 (outstanding aids)		
2. Quality of illustrations (print)	0 (not good)		1 (helpful)		2 (excellent)		E Total
3. Time and pacing	0 (bad)			1 (appropriate for broad range)			Grade
c. Recording answers	0 (complicated)		1 (standard)		2 (especially easy)		
3. Administrative Usability a. Administration 1. Test administration	0 (individual)		1 (small groups)		2 (large groups)		
2. Training of administrators	0 (psychometrist)			1 (school staff)			
3. Administration	0 (43+ minutes)			1 (42 minutes or less)			
b. Scoring	0 (subjective)		1 (difficult)		2 (simple)		
c. Interpretation 1. Norms a. Norm range	0 (restricted)			1 (broad)			
b. Score interpretation	0 (uncommon, abstruse)			1 (common, simple)			
c. Score conversion	0 (complicated)		1 (simple)		2 (clear, tables)		
d. Norm groups	0 (local, outdated, or poorly sampled)			1 (national, well sampled)			A Total
d. Score Interpreter	0 (psychometrist)			1 (school staff)			Grade
e. Can Decisions Be Made	0 doubtful		1 possible		2 probable	3 yes – charts and graphs	
4. Normed Technical Excellence a. Stability	not reported or less than .70 0		.70 to .80 1		.80 to .90 2	.90+ 3	
b. Internal Consistency	0		1		2	3	
c. Alternate form	0		1		2	3	
d. Replicability	0			1			N Total
e. Range of Coverage	0 no information		1 floor or ceiling reached		2 adequate	3 more than adequate	Grade
f. Scores	0 poorly graduated and uncommon		1 poorly graduated or uncommon		2 well graduated and standard		

Figure 1

tant is that of test format. The visual or auditory principles employed in test presentation were evaluated in terms of effective usage of Gestalt principles *(entry 5, Table 2)*. The evaluators looked for specific format features such as sufficiency of white space between items, visual or auditory coherence of item stems and alternatives, and effective use of colors as an aid in segregating items. The general quality of illustrations and print was also considered under physical format *(entry 6, Table 2)*.

For each scale, pacing or time limits were judged for their appropriateness for the subject matter and for the examinees *(entry 7, Table 2)*. Published statements regarding the speededness of tests were corroborated, when possible, by consulting item difficulty indexes and score distributions. In almost all cases, power was preferred to speed as an attribute of tests of educational output. The last aspect of appropriateness considered was the mode of response recording *(entry 8, Table 2)*. The more simple and direct connections between the item stem and the recording of a response were given more credit. All aspects of examinee appropriateness were rated relative to the specific grade level to which the test is directed.

Administrative Usability. After asking "What will it measure?" and "Is it designed for my students?", the next question was concerned with how usable the test is in terms of administration, scoring, interpretation, and decision making. These aspects of a test comprise the third criterion of the MEAN evaluations.

It was assumed that for general assessment of educational output, a test that can be administered to a large group is more desirable. Small group and individually administered tests were judged to be less usable for evaluation of instructional programs *(entry 9, Table 2)*; their usefulness for in-depth individual diagnosis was not in question. A second variable strongly affecting a test's utility is the training necessary to administer the test appropriately *(entry 10, Table 2)*. Since few schools have resident psychometrists and since most district psychometrists focus their attentions on individual student problems, a test was deemed to have greater utility if it could be administered by the school staff, preferably by the students' teacher. Tests were also credited if they fit into a typical class period and did not necessitate special scheduling *(entry 11, Table 2)*.

The utility of a test is further affected by the scoring procedure it requires *(entry 12, Table 2)*. Simple and objective hand or machine scoring of tests was considered optimal for utility; subjective scoring resulted in no credit. From a pragmatic viewpoint, while ease of administration and scoring are desirable, they are dwarfed by the importance of being able to interpret the scores and then of reaching some decision *(entry 18, Table 2)*. Tests from which prescriptive decisions can be made were given greater credit. Common, simple scores for interpretation earned a test more credit. In addition, a broad normative sample *(entry 13, Table 2)* which allows for both high and low achievement was rated superior to a restrictive sample; a current and representative norming sample was also rated higher *(entry 16, Table 2)*.

The normative score conversions were evaluated according to three criteria. If the derived scale is common and generally understood, the test was given more credit *(entry 14, Table 2)*. If the conversion is clear and unambiguous, the test earned credit over those with complicated, multi-stage conversions *(entry 15, Table 2)*. These two aspects of the derived scores determine in part who can interpret them. Tests yielding scores interpretable by school staff were preferred to those demanding the skills of a psychometrist *(entry 17, Table 2)*. The final pragmatic consideration of a test's utility rested on whether or not decisions, either individual or group, can be made on the basis of information in the test manuals.

Normed Technical Excellence. The last major criterion of the MEAN evaluation procedure was concerned with the reliability, replicability, and refinement of

Table 2
Mean Ratings of Tests on 24 Evaluative Criteria

Criteria		Range	Grade 1	Grade 3	Grade 5	Grade 6
Measurement Validity						
1.	Content and face validity	0-10	6.12	6.46	6.67	6.59
2.	Concurrent and predictive validity	0-5	1.00	0.96	1.14	1.26
Examinee Appropriateness						
3.	Content comprehension	0-4	3.14	3.12	3.22	3.16
4.	Instructions comprehension	0-4	3.21	3.22	3.26	3.20
5.	Visual principles of format	0-2	1.01	0.95	0.89	0.84
6.	Quality of illustrations	0-2	1.10	1.04	1.05	1.04
7.	Time and pacing	0-1	0.95	0.91	0.85	0.86
8.	Response recording	0-2	1.74	1.55	1.33	1.20
Administrative Usability						
9.	Test administration	0-2	1.11	1.47	1.65	1.80
10.	Training of administrators	0-1	0.75	0.81	0.87	0.94
11.	Administration	0-1	0.88	0.86	0.82	0.84
12.	Scoring	0-2	1.56	1.64	1.74	1.72
13.	Norm Range	0-1	0.69	0.74	0.82	0.76
14.	Score Interpretability	0-1	0.84	0.81	0.85	0.85
15.	Score conversion	0-2	1.34	1.41	1.44	1.36
16.	Norm representativeness	0-1	0.25	0.22	0.25	0.28
17.	Score interpreter	0-1	0.67	0.74	0.85	0.88
18.	Can decisions be made	0-3	1.32	1.39	1.46	1.43
Normed Technical Excellence						
19.	Test-retest reliability	0-3	0.15	0.23	0.25	0.24
20.	Internal-consistency	0-3	1.00	0.88	1.21	1.16
21.	Alternative form reliability	0-3	0.23	0.35	0.42	0.40
22.	Replicability	0-1	0.90	0.90	0.93	0.94
23.	Range of coverage	0-3	1.53	1.56	1.76	1.80
24.	Gradation of scores	0-2	1.46	1.38	1.58	1.57
Number of Instruments			318	380	477	508

measurement of the tests. Reliability was evaluated separately for published reports of test-retest *(entry 19, Table 2)*, internal-consistency *(entry 20, Table 2)*, and alternate-form estimates *(entry 21, Table 2)*. Closely related to the concept of test reliability is that of replicability of procedures to obtain the scores *(entry 22, Table 2)*. If procedures described in the test manual are complicated, subjective, and based upon abnormal samples, the test is clearly not replicable. Replicable procedures for obtaining scores were judged as more valuable.

The range of coverage is also an important aspect of a test's technical excellence. A broad developmental range which is appropriate for one level of assessment but which can also be applied to students above and below that level was preferred to a restricted range *(entry 23, Table 2)*. Related to the range problem is the refinement or gradation of the inter-individual comparison scores; the finer the gradation, the better the evaluation of the test *(entry 24, Table 2)*.

Each of the tests and scales, then, earned four scores; one for each of the MEAN criteria. These scores and their bases are published in *CSE Elementary School Test Evaluations*, by Hoepfner, Strickland, Stangel, Jansen, and Patalino (1970) in greater detail.* The four MEAN scores were, however, based upon twenty-four individual judgments. These discrete judgments were factor analyzed in order to uncover the characteristics of tests which actually do cohere. Table 2 presents the twenty-four criteria, the range of points possible for each of their evaluations, and the means of the consensual judgments for grades 1, 3, 5, and 6.

The separate judgments for each of the scales within each of the four grade levels were submitted to a principal-axes factor analysis. Initial solutions showed that only four factors appeared with regularity in all four grade levels. Because a fifth factor only appeared in two of the solutions (not chronologically adjacent grade levels), communality iterations were based on four factors. The matrices of intercorrelations among the rated characteristics are in *A Test of Tests*, CSE Report No. 69, by R. Hoepfner. The varimax factor loadings for the four factors and for the four grade levels are presented in Table 3.

Mean ratings of evaluative test qualities, as presented in Table 2, indicated no significant trends of increased or decreased quality over the four grade levels. One of the most salient findings in Table 2 is the relatively higher reliability estimate obtained through internal-consistency techniques. Whether or not this is an artifact of the ease of its estimation or the vulnerability of such estimates to extraneous inflationary factors cannot be determined.

It can also be seen from Table 2 that publishers provide very little evidence for the concurrent and predictive validities of their tests in the manuals they provide. This reflects, of course, the great costs to the publisher of such studies and the necessary delay from the time the manual is published to the time that various independent research findings can become incorporated into the publisher's documentation (if, indeed it ever is). Nonetheless, the typical rating on this criterion can be described as "very little evidence."

The comprehension levels of test items and instructions appear rather satisfactory, all means falling above the "probably appropriate" rating. This reflects the fact that most instruments at the elementary level are developed by curriculum experts at each grade level. Time and pacing and response recording procedures are also rated highly, probably for the same reason.

The visual principles and quality of illustrations for tests are rated at only slightly above average. Such mediocrity may be due to the expense of good graphics and layout or may be the result of a deliberate attempt by some publishers to avoid producing too polished a product (that might appear more commercial than educational).

The tests' major shortcomings in the area of Administrative Usability are the low quality of norm-group sampling and the failure to provide prescriptive decision rules on the basis of test results. Maintaining norm currency and obtaining national representativeness of the norm groups is the most expensive aspect of test publishing, and so it is not surprising that norms lack these qualities. Definitive and prescriptive decision rules

* A companion volume, *CSE-ECRC Preschool / Kindergarten Test Evaluations* (1971), treats early childhood tests in a similar manner.

Table 3
Varimax Factor Loadings for 24 Criteria for Four Grade Levels

Criteria	Grade 1				Grade 3				Grade 5				Grade 6			
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
1.	-.06	.13	.50	.01	-.02	.09	.56	.07	.02	.12	.69	.08	.11	.22	.64	-.02
2.	-.02	.01	.12	.73	.06	.27	.15	.63	.05	.17	-.11	.65	-.12	.20	-.13	.58
3.	-.23	-.08	.46	.03	-.29	-.16	.50	-.18	-.14	-.03	.54	-.04	-.05	-.04	.54	-.08
4.	-.45	.10	.14	.02	-.29	.16	.37	-.24	-.14	.21	.40	-.28	-.03	.12	.38	-.43
5.	-.51	-.05	.15	-.03	-.47	-.04	.15	-.67	-.33	-.08	.16	.11	-.39	.00	.32	.11
6.	-.29	-.02	.35	-.02	-.34	-.06	.14	.00	-.30	.03	.19	.05	-.29	.04	.26	.02
7.	-.20	-.12	.12	.07	-.19	-.06	.09	-.15	-.20	-.17	.02	-.03	-.19	-.23	.06	.14
8.	-.06	-.04	.07	.03	-.28	.06	.30	-.40	-.43	-.01	.21	-.14	-.20	.03	.25	-.36
9.	.73	.10	-.12	.00	.90	.07	.00	.03	.90	.18	.01	.14	.84	.11	.00	.17
10.	.91	-.01	.03	-.02	.88	-.03	-.07	.13	.84	.13	-.06	.16	.78	.05	-.04	.19
11.	.11	-.14	.01	-.12	.02	.04	.09	-.35	-.01	.00	.15	-.27	-.04	.01	-.03	-.15
12.	.52	.23	.05	.06	.55	.21	.11	-.03	.59	.18	.20	.06	.61	.34	.10	.04
13.	.00	.46	.07	.26	.02	.72	.17	-.03	.06	.58	.35	.02	.21	.65	.27	-.08
14.	.13	.39	.24	.20	.12	.48	.08	.18	.00	.60	.09	.11	-.08	.71	.02	.06
15.	.14	.25	.22	-.02	.05	.33	.32	-.05	.11	.34	.36	.03	.22	.59	.17	.13
16.	-.01	.18	.21	.43	.16	.36	.16	.43	.18	.17	.17	.55	.17	.28	.16	.40
17.	.84	.16	-.02	.02	.85	.08	-.05	.13	.91	.13	-.08	.13	.83	.06	.03	-.01
18.	.12	.20	.07	.20	.01	.17	.73	.20	-.03	.18	.74	.15	-.05	.27	.58	.19
19.	-.28	.09	.01	.37	.02	.09	.09	.53	.03	.03	-.01	.57	.00	.03	.03	.49
20.	.20	.38	.02	.50	.26	.49	.16	.32	.24	.51	.14	.22	.14	.58	.13	.02
21.	.14	.08	-.02	.35	.13	.24	-.03	.55	.08	.18	.10	.37	.03	.23	.08	.20
22.	.45	.01	.03	.25	.34	.18	.11	.03	.55	.03	.14	.12	.40	.29	-.03	.22
23.	.19	.53	-.05	.14	.15	.67	-.06	.22	.24	.65	.07	.15	.07	.72	.08	.05
24.	-.06	.97	.04	.06	.08	.82	-.03	.09	.09	.90	-.05	.12	.00	.83	-.04	.10

violate the often repeated (and frequently justified) warnings against too literal and decisive interpretations from faulty test scores. It seems that in following these well-intentioned warnings, the publishers make their instruments less useful for most educators who cannot operate with the ambiguous decision-making data provided for them.

While it is difficult to draw conclusions from the massive amounts of data provided in the correlation matrices, the outstanding finding is the relative lack of correlation between the ratings on the two kinds of test validity. The correlations between the ratings of face-content and concurrent-predictive validities range from -13 to +12, clearly demonstrating their independence, not only as constructs, but as results of actual practice in test construction and development.

The varimax solutions in Table 3 evidence considerable factorial constancy over the four grade levels. The fact that some instruments were common to more than one solution, being appropriate for a large grade span, cannot be hypothesized as accounting for this invariance, as there were few such overlapping instruments and the test evaluations were made separately at each grade level.

Factor A, consistently led by the variables of Test Administration, Training of Administrators, Score Interpreter, Scoring, and Replicability, clearly reflects a "Usability" dimension upon which tests can be placed. While not the same as the MEAN criterion of administrative usability, it is related as four of the eight variables having significant loadings are components of that criterion. It is interesting to note the consistent negative loadings for the Examinee Appropriateness ratings, especially for Visual Principles and Quality of Illustrations; perhaps this indicates that increased efforts to make tests usable have resulted in decreased attempts at making tests appropriate for the examinees.

Factor B is consistently led by the variables of Range of Coverage, Gradation of Scores, Norm Range, Score Interpretation, Score Conversion, and Internal-Consistency Reliability. This constellation of test attributes is named the "Norm Quality" factor, implying that normed tests tend to be good or bad in most of the norming attributes.

Factor C is led in all four grade levels by the variables of Ability to Make Decisions, Content and Construct Validity, and Content Comprehension. The factor probably reflects the amount of specificity of coverage of a test; tests being directed specifically to some focal goal area scored higher on these criteria. For this reason, Factor C is called the "Focus" factor.

Factor D is led by the variables of Concurrent and Predictive Validity, Norm Representatives, and Test-Retest Reliability. In several of the grade levels, the factor is further supported by the variables of Internal-Consistency and Alternate-Form reliabilities. This factor is parallel to Factor B and is called the "Psychometric Quality" factor. Apparently, publishers either exhaustively analyze their tests on most psychometric criteria, tend not to analyze on any of the criteria, or seek some consistent level of psychometric analysis.

Mean ratings of evaluations of tests, as presented in Table 2, indicate major shortcomings that characterize today's published instruments for elementary education. A factor analysis of these ratings revealed four consistent dimensions upon which tests actually vary: Usability, Norm Quality, Focus, and Psychometric Quality. The results of this analysis of tests should have many immediate and long-term implications for the improvement of assessment instrumentation by pointing out rather clearly some of the shortcomings that characterize today's published tests.