A TEST OF TESTS

by

Ralph Hoepfner

CSE Report No. 69
May 1971

School Evaluation Project
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

This report is a product of the School Evaluation Project. As part of the Center's Program on Evaluation of Educational Systems, the School Evaluation Project is designed to develop and field test sets of procedures which may be used by school evaluators and administrators engaged in evaluating schools--preschool, elementary, and secondary. The project is attempting to capitalize on the state of current knowledge to develop evaluation procedures which are appropriate especially to the first two stages of the Center's evaluation framework: Needs Assessment and Program Planning. The Center is concerned with developing procedures which will enable school principals and others to use information effectively in making valid decisions for improving student performance. The School Evaluation Project is currently field testing an evaluation KIT which is composed of a series of booklets describing how to conduct a needs assessment of an elementary school's student output.

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) the criteria may not be equally appropriate for all types of measures,

(b) the direct result 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 authors, 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.

PROCEDURE

The Center for the Study of Evaluation, in order to make an equable appraisal of the output measures published for use in evaluating elementary schools, programs, and students, developed a comprehensive objectives-based classification of needs-assessment areas for elementary education, and 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 were located. Each test or sub-scale was assigned to the pre-established

goal area into which it best fit. The tests 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 below.

Figure 1

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		

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. 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 exhaust the breadth of 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 1, page 9.)* 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 behaviors. A second aspect of measurement validity concerned the extent of reported empirical validation, either predictive or concurrent *(entry 2, Table 1)*.

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 1)*. 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 1)*. 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 important 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 1)*. 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 color as an aid in

segregating items. The general quality of illustrations and print was also considered under physical format *(entry 6, Table 1)*.

For each scale, pacing or time limits were judged for their appropriateness for the subject matter and for the examinees *(entry 7, Table 1)*. 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 1)*. 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 1)*; 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 1)*. 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 1)*.

The utility of a test is further affected by the scoring procedure it requires *(entry 12, Table 1)*. 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 1)*. 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 1)* 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 1)*.

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 1)*. If the conversion is clear and unambiguous, the test earned credit over those with complicated, multi-stage conversions *(entry 15, Table 1)*. 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 1)*. 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 measurement of the tests. Reliability was evaluated separately for published reports of test-retest *(entry 19, Table 1)*, internal-consistency *(entry 20, Table 1)*, and alternate-form estimates *(entry 21, Table 1)*. Closely related to the concept of test reliability is that of replicability of procedures to obtain the scores *(entry 22, Table 1)*. 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 restrictive range *(entry 23, Table 1)*. 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 1)*.

ANALYSIS

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 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 1 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 interrcorrelations among the rated characteristics are presented in Tables 2 through 5. The varimax factor loadings for the four factors and for the four grade levels are presented in Table 6.

RESULTS

Mean ratings of evaluative test qualities, as presented in Table 1, indicate no significant trends of increased or decreased quality over the four grade levels. One of the most salient findings in Table 1 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 1 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

Table 1

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

Table 2

Intercorrelations among 24 Ratings made on 318
Tests at the First Grade Level

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	28																							
2	09	54																						
3	25	11	28																					
4	13	04	28	23																				
5	07	02	10	23	28																			
6	20	07	28	11	32	21																		
7	-06	03	11	-01	19	08	07																	
8	-08	04	02	00	06	09	08	01																
9	-04	-03	-23	-32	-40	-22	-23	-08	55															
10	-11	00	-17	-42	-48	-17	-16	-04	68	83														
11	-04	-05	-01	-09	03	08	-04	11	07	10	04													
12	11	-02	-20	-20	-15	-20	-10	02	46	37	09	32												
13	16	25	-08	00	06	-02	00	16	12	-08	-05	22	28											
14	16	19	-06	-10	-02	08	01	12	-04	12	-01	24	37	26										
15	07	-08	21	06	-10	-11	-09	11	13	11	-09	17	19	11	13									
16	13	28	06	05	11	02	12	-04	-02	-04	-08	20	14	25	14	26								
17	-05	00	-23	-36	-43	-26	-18	-08	54	80	07	58	03	29	08	03	73							
18	40	23	24	06	-01	18	11	02	01	16	-10	08	15	30	21	25	12	54						
19	01	24	06	20	12	11	06	09	-23	-26	-13	-06	13	06	-04	26	-20	03	22					
20	06	36	03	-02	-21	-12	-06	-06	25	14	-18	21	31	20	15	33	24	22	17	43				
21	-02	36	-02	-04	-16	01	-11	-03	07	16	-01	03	08	10	00	13	11	10	01	21	15			
22	-06	16	-04	-16	-23	-18	-12	01	37	40	07	28	02	07	21	17	33	13	08	17	12	26		
23	00	12	-10	-04	-15	-02	-04	-14	18	23	-01	10	30	21	07	08	24	17	00	30	18	17	34	
24	11	07	-03	17	-04	06	-14	-06	04	-02	-15	16	39	39	25	20	10	23	14	37	12	-01	58	95

Note: Decimal points omitted; communality estimates in diagonal

Table 3

Intercorrelations among 24 Ratings made on 380 Tests at the Third-Grade Level

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	32																							
2	12	49																						
3	24	-07	39																					
4	18	-09	33	30																				
5	12	00	17	17	25																			
6	15	06	26	11	28	14																		
7	03	-09	10	06	20	05	06																	
8	03	-10	37	32	25	20	19	33																
9	-04	11	-29	-23	-40	-24	-19	-23	82															
10	-07	18	-27	-36	-47	-21	-15	-26	85	79														
11	00	-17	12	07	16	09	01	31	05	02	13													
12	14	03	-17	-07	-15	-28	-18	-14	56	36	07	36												
13	19	17	-06	16	-07	-02	01	08	02	-05	06	25	54											
14	13	33	-07	-05	-06	00	-03	-04	08	13	-04	11	39	28										
15	19	05	12	21	-06	-11	-05	11	07	-03	-04	14	30	17	22									
16	11	39	-13	-05	00	-08	-06	-16	19	13	-12	28	27	27	23	36								
17	01	16	-27	-28	-40	-29	-16	-31	74	80	01	54	05	27	-06	20	74							
18	55	27	29	27	05	01	04	12	01	-02	-01	07	26	14	32	25	-03	69						
19	-01	41	-03	-03	-01	02	-05	-11	07	06	-14	05	03	14	00	36	12	17	30					
20	13	36	-16	-01	-17	-05	00	-15	30	24	-13	29	38	33	16	43	30	33	22	44				
21	07	43	-17	-15	-17	-08	-25	-27	15	18	-12	08	19	15	10	30	16	12	36	27	38			
22	-02	15	-04	-03	-23	-22	-04	01	33	32	05	17	14	17	22	15	26	13	06	17	14	17		
23	05	36	-21	02	-16	-06	-06	-07	21	18	-01	14	49	32	16	31	19	10	18	39	36	24	52	
24	06	28	-18	10	-04	-06	-15	-02	16	09	03	19	53	46	27	31	16	15	12	49	23	09	62	69

Note: Decimal points omitted; communality estimates in diagonal.

Table 4

Intercorrelations among 24 Ratings made on 477
Tests at the Fifth-Grade Level

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	49																							
2	-13	46																						
3	34	-09	31																					
4	31	-21	30	30																				
5	08	00	18	08	15																			
6	10	08	28	12	21	13																		
7	-04	01	08	-01	24	09	07																	
8	12	-14	19	25	09	18	11	25																
9	01	17	-09	-10	-28	-20	-20	-40	87															
10	-02	21	-12	-16	-29	-17	-16	-40	85	76														
11	04	-23	19	12	-04	09	02	16	-02	-04	09													
12	17	06	01	02	-13	-20	-16	-25	62	42	03	42												
13	31	06	20	17	-09	04	-10	09	17	09	05	24	46											
14	08	16	08	10	01	04	-11	02	13	15	03	06	40	38										
15	29	04	13	15	05	-03	-07	00	10	07	00	27	35	34	26									
16	09	31	05	-13	01	-03	-15	-05	27	22	-17	26	23	19	15	39								
17	-03	14	-15	-16	-28	-28	-18	-43	85	86	-06	56	06	10	14	25	86							
18	58	07	34	32	11	14	00	10	01	-02	06	13	35	14	37	21	-07	60						
19	-10	40	02	-12	08	-02	-02	-10	13	11	-11	07	03	08	-02	32	12	11	32					
20	19	21	-05	09	-10	00	-17	-13	34	27	-14	29	34	32	22	28	31	27	10	38				
21	08	24	01	-02	-03	-02	-20	02	13	15	-01	09	11	17	11	37	10	11	23	21	18			
22	06	13	00	-06	-18	-18	-09	-18	51	52	01	36	09	05	19	13	47	13	08	17	13	33		
23	09	21	-02	12	-06	05	-16	-19	35	34	02	26	42	38	23	19	31	21	13	46	23	18	50	
24	08	26	-07	16	-06	-01	-14	-08	28	21	-05	23	51	56	29	21	24	13	12	48	17	06	63	83

Note: Decimal points omitted; communality estimates in diagonal.

Table 5

Intercorrelations among 24 Ratings made on 508
Tests at the Sixth Grade Level

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	47																							
2	-12	41																						
3	29	-14	30																					
4	24	-21	26	34																				
5	16	02	22	06	26																			
6	10	07	25	08	21	16																		
7	-05	03	05	-11	25	04	11																	
8	10	-16	17	37	07	13	-02	23																
9	08	04	-03	-06	-32	-20	-14	-28	75															
10	06	06	-04	-11	-34	-19	-10	-26	77	64														
11	-06	-14	09	-01	-08	04	00	10	00	-04	03													
12	23	-01	-02	03	-14	-17	-19	-10	58	38	-04	50												
13	35	01	18	16	-04	03	-17	14	23	17	00	37	55											
14	12	16	02	04	08	04	-12	04	02	03	04	14	49	51										
15	30	13	-01	03	04	-01	-12	-03	22	22	-03	38	42	50	44									
16	13	23	04	-13	05	-03	-12	-01	21	16	-17	28	31	21	28	29								
17	08	-12	-02	03	-31	-26	-15	-10	66	66	-09	57	24	-05	27	23	69							
18	52	14	26	22	18	17	03	04	06	03	-01	09	24	19	29	18	-11	45						
19	-03	33	03	-15	02	-01	06	-16	07	04	-01	01	02	05	04	31	09	10	24					
20	24	09	-06	10	-04	05	-14	03	18	08	-09	30	42	39	43	24	21	28	00	37				
21	12	16	04	-04	-02	00	-13	09	04	06	01	10	16	17	14	32	04	08	16	18	10			
22	12	13	-05	-14	-13	-12	-08	-19	40	42	-03	44	20	23	33	14	28	10	08	15	13	30		
23	18	17	02	13	04	07	-18	-08	18	13	06	29	51	45	42	17	08	28	05	45	21	26	54	
24	16	26	-05	09	-01	02	-13	-07	14	05	00	28	54	59	46	24	03	22	09	45	14	28	67	70

Note: Decimal points omitted; communality estimates in diagonal

Table 6

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	-.07	-.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	-.13	-.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	.67	.20	.01	.17	.78	.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	.02	.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

become incorporated into the publishers 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 appears 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 principals 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 suprising that norms lack these qualities. Definitive and prescriptive decision rules 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 (Tables 2 through 5), 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 6 evidence considerable factorial invariance 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 this criterion. It is interesting to note the consistent negative loadings for the Examinee Appropriateness ratings, especially for Visual Principals 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 Representativeness, 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 a parallel to Factor B and is called the "Psychometric Quality" factor. Apparently, publishers either exhaustively analyze their tests on all psychometric criteria, tend not to analyze on any of the criteria, or seek some consistent level of psychometric analysis.

CONCLUSIONS

Mean ratings of evaluations of tests, as presented in Table 1, 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 test of tests should have many immediate and long-term implications for the improvement of assessment instrumentation by pointing out rather clearly the shortcomings that characterize today's published tests.

References

Hoepfner, R., Strickland, G. P., Stangel, G., Jansen, P., & Patalino, M. CSE Elementary School Test Evaluations. Los Angeles: University of California, Los Angeles, 1970.