WHAT TEST PUBLISHERS PUBLISH

By

Ralph Hoepfner and William J. Doherty

CSE Report No. 75

February 1972

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

Attempts to systematically evaluate educational materials (e.g., EPIE Educational Product Reports and CSE Test Evaluations) have as their principal goal the provision of evaluative information to the user of the materials. A not unexpected side benefit results when the producers of the materials are convinced of the adequacy and relevance of the evaluations and the criteria upon which they were made, and execute product improvements in efforts to meet those criteria more fully. Materials evaluations might also be expected to reflect upon the goals and priorities held by the producers. Provided that the producers are well-intentioned and have educational improvement in mind to some degree, their development priorities should be evident from summaries of the evaluations of their individual products.

Such a systematic and exhaustive evaluation of tests for elementary school has been completed (Hoepfner, Strickland, Stangel, Jansen, and Patalino, 1970), and the results of it were considered valuable source data for discovering the peculiar strengths and weaknesses of the test producers. Evaluative ratings were available for some 1,600 different published scales appropriate for grades 1, 3, 5, and 6. The ratings had been performed by staff members of the Center for the Study of Evaluation who were experienced in elementary schools or with the technical aspects of testing. They arrived at the ratings through the implementation of the MEAN test evaluation system.

The MEAN test evaluation system comprises ratings on 24 different aspects of standardized published tests. The 24 aspects are grouped into four major categories: Measurement Validity, Examinee Appropriateness, Administrative Usability, and Normed Technical Excellence. Within any one of the four categories, a test can earn up to 15 points. The rating form and a detailed description of the rationale and implementation of the MEAN system can be found in Hoepfner, et al. (1970). A brief description of each of the MEAN categories is included here for reference:

Measurement Validity. Ratings on this criterion reflect how well the test measures the specific educational goal for which it was developed or to which it was assigned. In addition, weight was given to the amount of evidence of criterion-related validity which was supplied by the publisher in the test manual.

Examinee Appropriateness. The second category comprises ratings of the appropriateness of such aspects as the comprehension level of the test and the instructions, the physical format, and the required response mode; each rating being made in reference to the grade level for which the test was designed or suggested.

Administrative Usability. Ratings in this category reflect how usable the test is in terms of administration, scoring, interpretation, and educational decision making.

Normed Technical Excellence. The last criterion concentrates on the technical or measurement characteristics of the test. The specific ratings that constitute this category were directed toward the test's reliability, replicability, and refinement of measurement.

METHOD

The 1600 tests and scales rated were from 39 different publishers and distributors. Characterizations of publishers represented by only a few scales would expectably lead to unreliable conclusions regarding their priorities. In order to circumvent this problem, a tally of the number of scales for each test publisher was made. The publishers and the number of test evaluated are listed in Table 1.

Several selection criteria were then applied to the list of publishers to obtain the final sample of publishers, each with a sufficient number of scales represented to afford a reliable characterization. The criteria for selection were concerned with the number of tests that represented the publisher and the number of scales and tests at each grade level. Specifically, four criteria were employed in the selection procedure. First, for selection the test publisher had to have fifty or more scales rated (see Table 1) within the four grade levels. Second, the scales had to be derived from at least five different tests or test batteries. Third, at any particular grade level, the test publisher had to have at least ten different scales rated. As a final criterion, it was decided that only test publishers would be examined and that all test distributors would be eliminated from further consideration. The rationale for this was that test distributors generally have little control over or effect on the characteristics of the test that they distribute. It was planned that utilization of these criteria would result in the inclusion of only those test publishers who could be considered "major" publishers and in the minimization of dependencies within the characterizations caused by the separate ratings of different scales of the same test or battery (i.e. within a large battery, many subscale characteristics are common to all subscales; but may not really characterize the priorities of the publisher). Applying these criteria to the publishers listed in Table 1 resulted in the retention of seven publishers at grades 3, 5 and 6; and of six publishers at grade 1. The publishers whose tests were to be further analyzed were: BMC, CTB, HBJ, HMC, PC, STS, and SRA; the first was not included at grade 1, as only four tests were rated.

Table 1

Thirty-Nine Publishers and the Numbers of their Elementary-Level Scales

Code	Publisher	Number of Scales
AAJE	American Association for Jewish Education	2
AGS	American Guidance Service, Inc.	33
AP	Association Press	15
BMC	Bobbs-Merrill Co., Inc.	56
BEM	Bureau of Educational Measurements	20
BERS	Bureau of Educational Research and Service	6
CTB	CTB/McGraw-Hill	335
CPS	Center for Psychological Service	2
CDRT	Committee on Diagnostic Reading Tests	2
EDL	Educational Development Laboratories, Inc.	4
EITS	Educational and Industrial Testing Service	58
ETS	Educational Testing Service	45
EETSA	Educator's-Employer's Tests and Services Association	2
EPS	Educator's Publishing Service	15
FPC	Follett Publishing Company	12
GA	Guidance Associates	12
GC	Ginn and Company	70
GTA	Guidance Testing Associates	24
HBJ	Harcourt, Brace, Javanovich	238
HMC	Houghton-Mifflin Company	140
IPAT	Institute for Personality and Ability Testing	30
LRA	Language Research Associates	2
LC	Lyons and Carnahan	22
M	Monitor	1
OBL	Oliver and Boyd, Ltd.	2
PC	Psychological Corporation	112
PTS	Psychological Test Specialists	8
PA	Psychometric Affiliates	18
RGS	Robert Gibson and Sons, Ltd.	6
STS	Scholastic Testing Services, Inc.	98
SRA	Science Research Associates	126
SEP	Slosson Educational Publications	8
SMP	St. Martin's Press, Inc.	4
SVC	Steck-Vaughn Company	31
TCP	Teacher's College Press	57
UIP	University of Illinois Press	39
WPS	Western Psychological Services	22
WLRF	Winter Haven Lions Research Foundation	1
ZBC	Zaner-Bloser Company	4
	Total Number of Tests	1,683

STATISTICAL ANALYSIS

To determine whether differences existed between the test publishers in terms of their test ratings, and particularly if differences in profiles (i.e. the ratings on all four of the MEAN criteria) existed, an analysis of variance with repeated measures was selected. Specifically, at grades 3, 5, and 6, a 7x4 analysis of variance with repeated measures was performed. At grade 1 a 6x4 analysis was performed. The computational procedure as given by Winer (1962) was programmed and the computations were performed by an IBM 360/91 computer.

Following the analyses of variance, a cluster analysis was performed in lieu of individual contrasts, primarily because the aim of this study was to describe test publishers' characteristics in terms of their tests rather than to determine that one test publisher might be "better" than another. The cluster analysis was performed by submitting the mean values for each publisher on each of the four MEAN categories to the BMDP2M program (Dixon, 1970). The means as well as the sample sizes upon which they are based can be found in Table 2. The cluster-analysis program computes an initial matrix of distances between the original cases, the distances being the square root of the sums of the squares of differences. The program then clusters the two cases having the closest distances and treats them as one case, then recomputes the distances. This procedure is iterated until one total group is achieved. The initial distance matrices for the four grade levels may be found in Table 3.

Table 2

Test Sample Sizes and Average MEAN Ratings for Seven Test Publishers over Four Grade Levels

Publisher	Grade 1	Grade 3	Grade 5	Grade 6

BMC	*	n= 17
		M= 6.47
		E=10.35
		A=11.18
		N= 5.06
			n= 19
			M= 6.94
			E= 9.95
			A=11.42
			N= 6.32
				n= 16
				M= 6.88
				E=10.00
				A=11.50
				N= 6.75

CTB	n= 51
	M= 6.67
	E=10.18
	A=10.94
	N= 4.57
		n= 73
		M= 6.96
		E=10.77
		A=11.29
		N= 5.68
			n=103
			M= 7.47
			E=10.91
			A=11.74
			N= 6.21
				n= 109
				M= 7.39
				E=10.86
				A=11.75
				N= 5.98

HBJ	n= 37
	M= 8.16
	E=11.35
	A=10.16
	N= 6.11
		n= 55
		M= 8.24
		E=10.56
		A=11.07
		N= 6.09
			n= 73
			M= 8.22
			E=10.43
			A=11.60
			N= 6.32
				n= 71
				M= 8.23
				E=10.48
				A=11.49
				N= 6.09

HMC	n= 31
	M= 6.71
	E=11.19
	A=10.35
	N= 5.16
		n= 40
		M= 8.83
		E= 9.48
		A=11.92
		N= 7.64
			n= 34
			M= 9.38
			E= 8.94
			A=12.62
			N= 8.68
				n= 35
				M= 9.40
				E= 9.03
				A=12.66
				N= 8.60

PC	n= 32
	M= 7.44
	E=12.56
	A= 6.09
	N= 6.41
		n= 20
		M= 7.10
		E=12.45
		A= 5.45
		N= 5.20
			n= 31
			M= 7.97
			E=12.13
			A= 6.94
			N= 4.97
				n= 30
				M= 7.93
				E=12.67
				A= 6.90
				N= 4.97

STS	n= 18
	M= 7.56
	E=10.28
	A=11.56
	N= 5.61
		n= 20
		M= 6.70
		E=10.80
		A=10.80
		N= 4.15
			n= 29
			M= 7.17
			E=10.83
			A=11.83
			N= 6.03
				n= 28
				M= 7.29
				E=10.75
				A=11.89
				N= 6.21

SRA	n= 20
	M= 7.40
	E=11.40
	A=12.50
	N= 7.20
		n= 17
		M= 8.53
		E=11.82
		A=12.88
		N= 7.00
			n= 30
			M= 8.60
			E=10.47
			A=13.23
			N= 8.30
				n= 57
				M= 8.67
				E= 9.58
				A=12.86
				N= 7.91

*Number of tests insufficient to support reliable means.

Table 3

Euclidean Distances Between Publishers at Four Grade Levels

Grade 1

	CTB	HMC	SRA	STS	HBJ	PC
CTB		1.35	3.48	1.96	3.42	4.24
HMC			2.69	1.97	2.78	3.10
SRA				2.19	2.08	3.30
STS					1.83	3.70
HBJ						2.66
PC						

Grade 3

	BMC	CTB	STS	HMC	SRA	HBJ	PC
BMC		0.85	0.93	3.41	3.17	2.05	3.28
CTB			1.33	2.89	2.35	1.40	3.01
STS				3.94	3.34	2.30	2.96
HMC					2.52	1.86	4.90
SRA						1.70	3.81
HBJ							3.34
PC							

Grade 5

	BMC	STS	CTB	HBJ	SRA	HMC	PC
BMC		0.99	1.18	1.57	2.66	3.57	3.51
STS			0.38	1.31	2.52	3.82	3.01
CTB				1.02	2.24	3.58	2.88
HBJ					1.76	2.76	3.07
SRA						1.86	4.40
HMC							5.39
PC							

Grade 6

	BMC	STS	CTB	HBJ	SRA	HMC	PC
BMC		0.93	1.13	1.68	2.38	3.40	3.74
STS			0.25	1.12	2.37	3.45	3.23
CTB				1.03	2.47	3.53	3.05
HBJ					1.86	2.79	3.11
SRA						1.12	4.71
HMC							5.42
PC							

RESULTS

The results of the analyses of variance for the four grades are shown in Table 4. At each grade level the results were highly significant, not only for the main effects (publisher and criterion) but also for the interaction. It can be seen from Table 4 that a definite difference exists among the publishers in their average MEAN ratings. The F-value at all four grade levels for this publisher effect is significant beyond the .01 level. The other main effect, the differences between the various categories of the MEAN rating system, is also highly significant at all four grade levels. This second difference will not be pursued, however, as it is not really relevant to the aims of this study. The significant tests which were of primary concern in the study involved the interaction terms. Once again, in all four analyses, the F-ratios were significant beyond the .01 level. This result thus fortified our belief that it would be possible to characterize groups of publishers in terms of average MEAN rating profiles. The figural representation of the average ratings is presented in Figure 1.

From the figures, it can clearly be seen why a significant interaction effect was found. What is even more interesting is that groups of similar profile types seem to emerge from the figure. For example, at the third-grade level, the profiles of the publishers CTB, SRA, and STS are very similar in shape although different in level. Similarly, the profiles of HBJ and HMC are similar in shape, while differing slightly in level. This type of "eyeball" clustering, while perhaps interesting, can only be conjectural unless supported by more tangible evidence. It was for this reason that the cluster analyses were performed.

Table 4

Analyses of Variance over Publishers and Rating Criteria for Four Grade Levels*

Grade 1

Source	SS	df	MS	F
Between Publishers	189.62	5	37.92	8.04
Scales within Publishers	863.63	183	4.72	
Between Rating Criteria	3,593.57	3	1,197.86	471.80
Publisher X Rating Criteria Interaction	845.56	15	56.37	22.20
Scales by Rating Criteria within Publishers	1,393.88	549	2.54	

Grade 3

Source	SS	df	MS	F
Between Publishers	374.13	6	62.35	9.03
Scales within Publishers	1,623.06	235	6.91	
Between Rating Criteria	4,216.13	3	1,405.38	397.05
Publisher X Rating Criteria Interaction	888.25	18	49.35	13.94
Scales by Rating Criteria within Publishers	2,495.38	705	3.54	

Grade 5

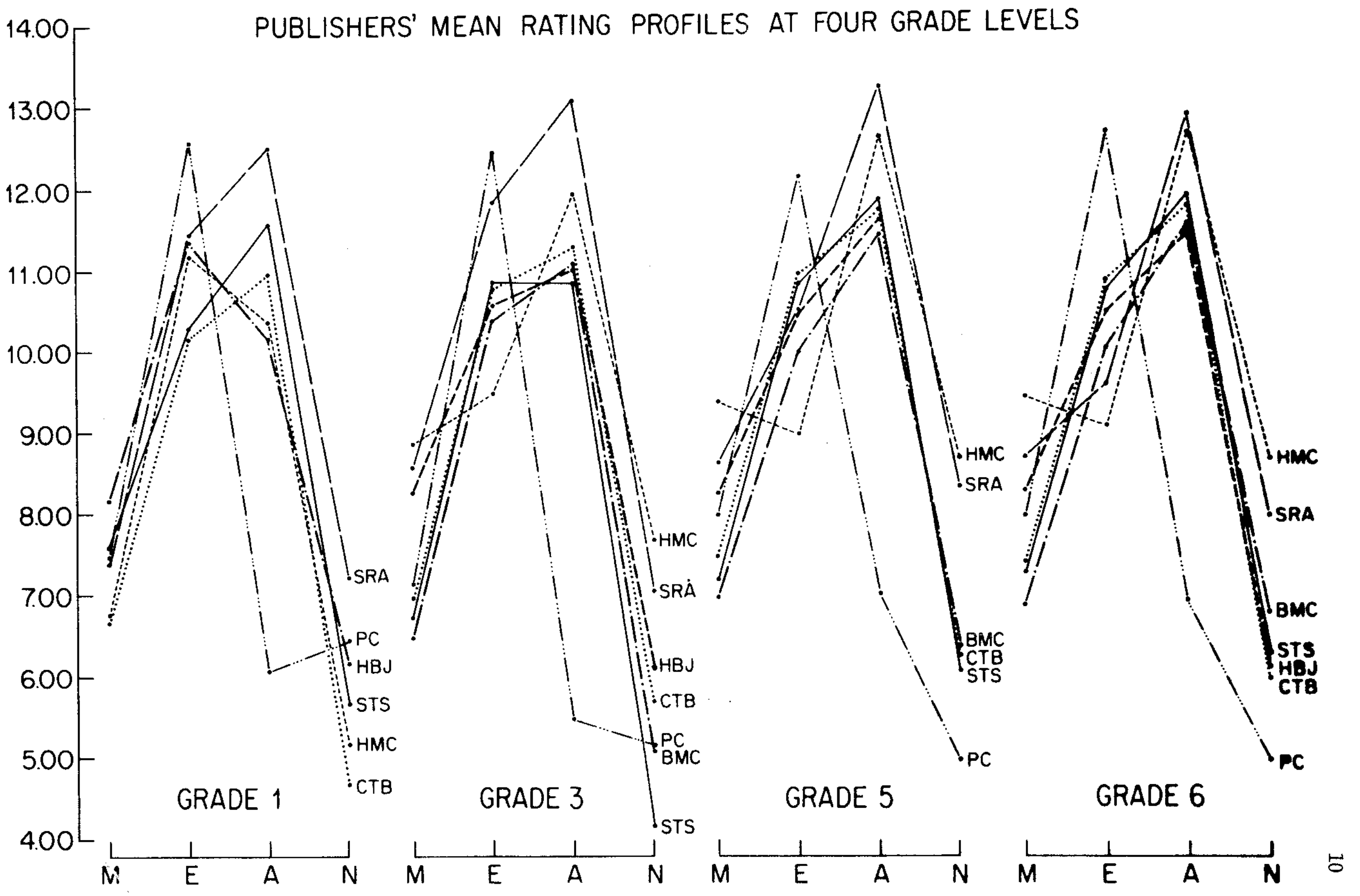
Source	SS	df	MS	F
Between Publishers	385.56	6	64.26	9.30
Scales within Publishers	2,156.00	312	6.91	
Between Rating Criteria	4,979.94	3	1,659.98	600.77
Publisher X Rating Criteria Interaction	1,086.31	18	60.35	21.84
Scales by Rating Criteria within Publishers	2,586.25	936	2.76	

Grade 6

Source	SS	df	MS	F
Between Publishers	356.63	6	59.44	9.14
Scales within Publishers	2,203.94	339	6.50	
Between Rating Criteria	5,375.81	3	1,791.94	680.22
Publisher X Rating Criteria Interaction	1,212.81	18	67.38	25.58
Scales by Rating Criteria within Publishers	2,679.13	1,017	2.63	

*Two error terms necessitated by (two-factor with repeated measures on one factor) design (Winer, 1962, pp 302-310).

Figure 1


PUBLISHERS' MEAN RATING PROFILES AT FOUR GRADE LEVELS
14.00
13.00
12.00
11.00
10.00
9.00
8.00
7.00
6.00
5.00
4.00
GRADE 1
SRA
PC
HBJ
STS
HMC
CTB
GRADE 3
HMC
SRA
HBJ
CTB
PC
BMC
STS
GRADE 5
HMC
SRA
BMC
CTB
STS
PC
GRADE 6
HMC
SRA
BMC
STS
HBJ
CTB
PC
M
E
A
N
M
E
A
N
M
E
A
N
M
E
A
N

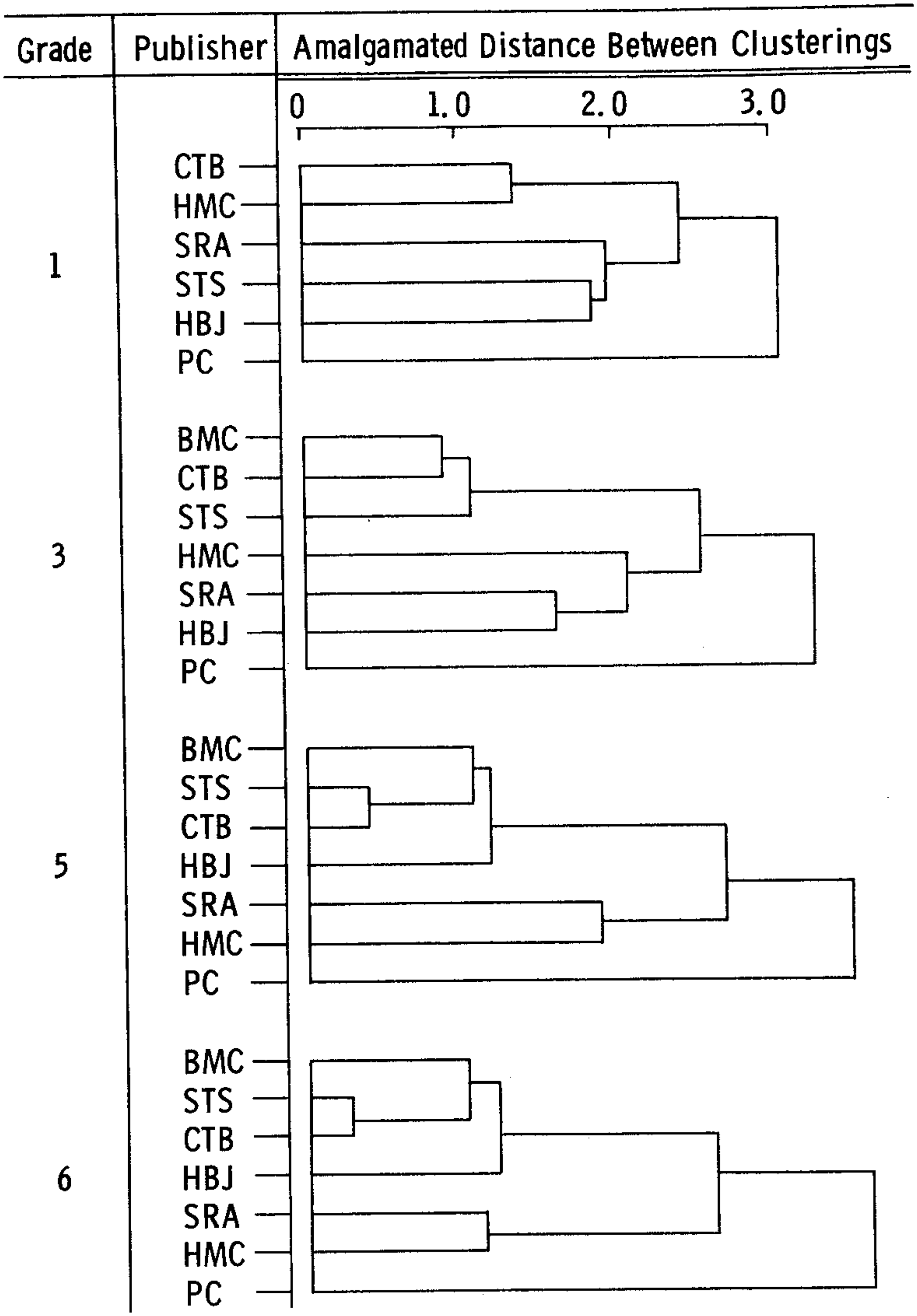
For the cluster analyses, the Euclidean distance was employed (see Table 3) following standardization of the various MEAN categories. It was felt that this approach was appropriate since the Euclidean distance measure is common and the standardization removes artifactual scale differences between the categories over which the clustering is to occur. The results of the clustering for the four grade levels are graphically shown in Figure 2.

Although there is some variation in the orders of groupings, it seems that some fairly stable clusters of publishers emerge over the four grade levels. One such cluster contains the publishers BMC, CTB, and STS, with HBJ being a sort of borderline member. The second cluster is made up of HMC and SRA, with HBJ being a borderline member once again. The third cluster is defined by PC alone. Referring to Figure 1, it can be seen rather quickly how the clusters have come about. The publishers in the first cluster have very similar average values on the middle two MEAN categories (E and A) over all four grades. In addition, at grades 5 and 6, the average ratings for the last MEAN category are extremely close for these publishers. At grades 1 and 3, it is the first MEAN category (M) upon which they are similar. The result is that the shapes and levels of the profiles for these publishers are very similar, and in terms of the summary test evaluations (see Hoepfner, et al., 1970) the publishers can be characterized as producing tests that are good in terms of administrative usability; fair for examinee appropriateness; but poor for measurement validity and normed technical excellence.

The second cluster of HMC and SRA have similar ratings on three of the MEAN categories over grades 3, 5, and 6. For the categories M, A, and

Figure 2

CLUSTER DIAGRAMS OF PUBLISHERS AT FOUR GRADE LEVELS



N the two publishers are very close, but they are disparate on the E category. At grade 1, the pattern is slightly different, with the similarity being only on the M and E categories and disparities on the A and N categories. In general, however, this cluster could be described as producing tests good in administrative usability, fair in measurement validity and examinee appropriateness, and poor in normed technical excellence.

The last cluster is defined solely by PC. This publisher has a MEAN profile totally distinctive from the other publishers. In all cases, the profile was the same, starting at about an average rating for the M criterion; going very high on the E category; and then dropping sharply to low values on the A and N categories. It would seem that this publisher has concentrated its major efforts in the area of examinee appropriateness, partly through the heavy usage of individually-administered tests (which, incidentally, partly accounts for the low administrative usability rating). The tests for this publisher are characterized as being good on examinee appropriateness, but poor on measurement validity, administrative usability, and normed technical excellence.

SUMMARY

While there is a great temptation for the producers and possessors of systematic and objective evaluative information on institutions and products to wield the power that information affords in order to effect some favored course of action or policy (indeed, there was the temptation to entitle this paper "Publishers Perish!"), the intent of this report is to present descriptive information regarding the various test publishers' priorities. The data support the supposition that the test publishers do differ in their priorities as reflected in their test's characteristics. In addition to the observed differences, three clusters of test publishers appeared. Cluster I publishers can be characterized as producing tests highly usable administratively and fairly good in terms of examinee appropriateness. Cluster II is like the first cluster, but its publishers produce tests with greater relevance and validity. Cluster III, with only one publisher, has emphasized the examinee appropriateness of its tests to the neglect of other test qualities. Although the obtained publisher profiles and clusters could be used as rough guides by test purchasers, a far more important utilization would be as the guidelines for self-improvement of the publishers through improving their products' qualities.

References

Dixon, W. J. (Ed.) BMD Biomedical Computer Programs. Berkeley: University of California Press, 1970.

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

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1962.