The Frequency and Quality of Measures Utilized in Federally-Sponsored Research on Children and Adolescents

Author(s): Stephen P. Heyneman and Pamela Cope Mintz

Source: *American Educational Research Journal*, Vol. 14, No. 2 (Spring, 1977), pp. 99-113
Published by: American Educational Research Association
Stable URL: http://www.jstor.org/stable/1162703
Accessed: 12-11-2017 21:12 UTC

American Educational Research Journal
Spring 1977, Vol. 14, No. 2, Pp. 99–113

The Frequency and Quality of Measures Utilized in Federally-Sponsored Research on Children and Adolescents

STEPHEN P. HEYNEMAN
The World Bank

PAMELA COPE MINTZ
George Washington University

Every test intended for use in FY '75 Federally-funded research on children or youth was placed on a list. The list eventually comprised the titles of 1,570 instruments. Some were mentioned by many principal investigators in their proposals; others were mentioned by only one. Some tests were highly respected instruments; others were not. The question pursued was whether there was a relationship between an instrument's quality and the frequency with which it was used. To gain a sense of a measure's quality, we utilized the numerical ratings published by the UCLA Center for the Study of Evaluation. For frequency of use we counted the number of times an instrument was mentioned in 3,538 research proposals on children or youth which are currently being sponsored by the Federal government.

There is a positive relationship between the quality of tests and their frequency of use. But the degree is not equally strong from one test category to another. A preference for the better rated instruments is particularly evident with tests of academic achievement. More equivocal results appear with respect to tests of vocational skills and intelligence, though in certain respects researchers are definitely using the better of those available in these two categories. The anomaly lies in the categories of reading tests and tests of personality, where the higher rated of the tests have no better chance of being utilized than those judged to be of poor quality. All sponsored researchers need not use the same instruments. But the fact that there are particular subject areas such as in reading

and personality where the higher rated tests are less likely to be used points to the need for special attention when choosing instruments for future research.

Each project proposal funded by 23 Federal agencies during FY '75 provided data for an analysis under the direction of the Interagency Panels for Research and Development on Early Childhood and for Research and Development on Adolescence. The information garnered from the 3,538 proposals formed the basis for the Third Annual Report on adolescence research and the Fifth Annual Report on early childhood research (Hertz & Mann, 1975; Heyneman, 1975). These reports discussed in some detail the patterns of Federal interests in subjects such as cognitive development, physical handicaps, career education, and day care.

However, in addition to noting the subject matter of each funded proposal, the coders working at the agencies were requested to note the title of any instrument mentioned in the proposal which was intended for use as a test or measure. No effort was made to exclude titles which were developed for use only in one particular project—such as attitudinal questionnaires or tests under construction. Consequently, if it had a "name," it was recorded. This paper is a review and brief analysis of this list.[1]

From these proposals 1,570 instrument names were noted and alphabetized. The first task was to discover which among them were referenced in the test bibliographic literature, to determine which titles were being utilized by particular research projects and where one might turn for a published description. We gathered 33 different test bibliographies and began the process of looking up each title until we came upon a reference for it. The more widely known and comprehensive bibliographies were consulted first: e.g., Buros (1974), Chun, Cobb and French (1975), Robinson and Shaver (1969). If a title was found in one, its volume and page number were noted for future reference, and we then went on to the next test title. Thus, we have not located every citation of every test title, but have differentiated those with at least one reference citation from those with none.

One peculiarity of the data from the proposals is that the official names for tests are not always used, and it was not rare to get three, four or even five different titles for the same test. Some judgment was required to identify tests by an acronym, with a word or two missing, or with an author's name misspelled. Each title was looked up by both the test name and the author's name.

In this way, we created two initial lists: one which we labeled as *referenced* and a second labeled *non-referenced.* In all, 1086 titles were put on our non-referenced list. Most were mentioned in only one project proposal, having,

1. The list of both referenced tests and non-referenced tests, and a complete bibliography of test references used, may be obtained from the Interagency Panels, the Social Research Group, or the ERIC Reproduction Service No. ED129-905.

therefore, a frequency of one. It is presumed that many of these measures were being developed by the project's principal investigator and may appear in future test bibliographies.

On the other list went the 484 titles which were located in one of the reference bibliographies. With respect to this referenced list, our first task was to acquire some indication of their purpose. For this, we adopted the classification scheme utilized by Buros (1974) in his book *Tests In Print, II* which includes over one hundred categories of test subjects (intelligence, achievement, attitude, etc.).

FREQUENCY OF USE

Out of the 484 referenced test titles, 203 (42 percent) were measures of character and personality; 64 (13 percent) were measures of intelligence; and 65 (13 percent) were measures of reading ability. Academic achievement batteries, tests of vocational skills, speech and hearing problems, and sensory-motor abilities accounted for approximately 25 (5 percent) each. These figures are displayed for each test subject category in Table 1.

However, having a large number of test titles in a subject category does not necessarily indicate that the tests in that category were utilized more often. For example, the few academic achievement batteries (5 percent of the titles) accounted for over a fourth of all tests and measures used (26 percent) (Table 2), and intelligence test titles (13 percent) accounted for almost a third of test usage (32 percent). Conversely, though the 203 character and personality test titles made up 42 percent, their application accounted for half of that, only 21 percent of the frequency. In sum, though there were fewer academic achievement and intelligence test titles than there were character and personality titles, titles in the former two categories tend to be used more frequently.

RATINGS OF TEST QUALITY

Though informed and helpful opinion concerning tests could be found in all of the bibliographies used, we have chosen the ordinal ratings of test "quality" which have been published by the Center for the Study of Evaluation (CSE) at UCLA (Hoepfner, 1970; Hoepfner, Stern, & Nummendal, 1971; Hoepfner, Hemenway, DeMuth, Tenopyr, Granville, Petrosko, Krakower, Silberstein, & Nadeau, 1972; Hoepfner, Conniff, Petrosko, Watkins, Erlich, Todaro, & Hoyt, 1974a; Hoepfner, Conniff, McGuire, Klibanoff, Stangel, Lee, & Rest, 1974b; Hoepfner, Conniff, Hufano, Bastone, Ogilvie, Hunter, & Johnson, 1974c). To the best of our knowledge, this was the only source which has attempted to evaluate tests by rating them numerically.

Of the four categories of quality specified by CSE, we have elected to consider the two entitled (A) *Measurement Validity,* and (B) *Normed Technical Excellence.* The other two categories (C) *Examinee Appropriateness* and (D) *Administrative Usability,* were omitted because they were designed to help the classroom teacher or educational administrator. Since test ratings varied in their appropriateness between preschool and high school, and since our population of government

TABLE 1
Number and Percentage of Referenced Test Titles by Subject Category

Subject Category[a]	Percentage		(N=484) (N)	
Academic Achievement Batteries	5.4		(26)	
Character and Personality				
General	19.2		(93)	
Non-Projective	18.4		(89)	
Projective	4.3		(21)	
Total		41.9		(203)
English				
General	1.2		(6)	
Spelling	.4		(2)	
Vocabulary	.2		(1)	
Music	.4		(2)	
Foreign Languages				
General	.2		(1)	
English	.8		(4)	
French	.2		(1)	
Spanish	.2		(1)	
Intelligence				
General	.2		(1)	
Group	5.2		(25)	
Individual	6.4		(31)	
Specific	1.5		(7)	
Total		13.2		(64)
Mathematics				
General	.2		(1)	
Algebra	.2		(1)	
Arithmetic	1.0		(5)	
Computational and Scoring Devices	.2		(1)	
Education	.2		(1)	
Industrial Arts	.2		(1)	
Listening Comprehension	.4		(2)	
Psychology	.2		(1)	

TABLE 1 (*Continued*)

Subject Category[a]	Percentage		(N=484) (N)	
Socioeconomic Status	1.0		(5)	
Multi Aptitude Batteries	1.5		(7)	
Reading				
General	2.7		(13)	
Diagnostic	2.7		(13)	
Miscellaneous	.8		(4)	
Oral	.6		(3)	
Readiness	5.2		(25)	
Speed	.2		(1)	
Study Skills	1.2		(6)	
Total		13.4		(65)
General Science	.2		(1)	
Biology	.2		(1)	
Sensory-Motor				
General	4.1		(20)	
Motor	.6		(3)	
Vision	.6		(3)	
Total		5.4		(26)
Vocational Tests				
General	1.7		(8)	
Clerical	.2		(1)	
Miscellaneous	.2		(1)	
Interest Inventory	1.9		(9)	
Manual Dexterity	.6		(3)	
Mechanical Ability	.6		(3)	
Total		5.2		(25)
Speech and Hearing				
General	.2		(1)	
Hearing	1.2		(6)	
Speech	8.9		(19)	
Total		5.4		(26)
Learning Disabilities	.8		(4)	

[a]These categories are identical to those defined in Buros (1974).

TABLE 2
Number and Percentage of Referenced Test Use Frequencies by Subject Category

Subject Category[a]	Percentage		(N=2721) (N)	
Academic Achievement Batteries	26.2		(714)	
Character and Personality				
General	9.0		(244)	
Non-Projective	7.7		(210)	
Projective	4.3		(118)	
Total		21.0		(572)
English				
General	.2		(5)	
Spelling	.1		(2)	
Vocabulary	—		(1)	
Music	.1		(2)	
Foreign Languages				
General	.1		(3)	
English	.3		(9)	
French	—		(1)	
Spanish	.1		(2)	
Intelligence				
General	.1		(2)	
Group	9.7		(265)	
Individual	19.9		(540)	
Specific	2.0		(54)	
Total		31.6		(861)
Mathematics				
General	.2		(4)	
Algebra	—		(1)	
Arithmetic	.6		(16)	
Computational and Scoring Devices	.1		(3)	
Education	—		(1)	
Industrial Arts	—		(1)	
Listening Comprehension	.3		(7)	
Psychology	.0		(0)	

TABLE 2 (*Continued*)

Subject Category[a]	Percentage		(N=2721) (N)	
Socioeconomic Status	.2		(5)	
Multi-Aptitude Batteries	.7		(19)	
Reading				
General	2.7		(72)	
Diagnostic	1.8		(49)	
Miscellaneous	.9		(25)	
Oral	.8		(22)	
Readiness	4.0		(108)	
Speed	—		(1)	
Study Skills	.3		(9)	
Total		10.5		(286)
General Science	—		(1)	
Biology	—		(1)	
Sensory-Motor				
General	2.9		(80)	
Motor	.1		(3)	
Vision	.2		(4)	
Total		3.2		(87)
Vocational Tests				
General	.6		(17)	
Clerical	—		(1)	
Miscellaneous	—		(1)	
Interest Investory	.7		(19)	
Manual Dexterity	.1		(3)	
Mechanical Ability	.1		(3)	
Total		1.7		(45)
Speech and Hearing				
General	.2		(5)	
Hearing	.2		(4)	
Speech	2.1		(57)	
Total		2.4		(66)
Learning Disabilities	.2		(6)	

[a]These categories are identical to those defined in Buros (1974).

research principal investigators would have to make these decisions for their own particular range of subjects under study, we decided to use only the Validity and Technical Excellence categories. Based upon universally acknowledged principles of empirical testing, these latter categories seemed more apropos for gauging the tests and measures used in adolescence and early childhood research, since these two subjects are not limited to education, but span the full gamut of the social and biological sciences.

With respect to Measurement Validity, CSE usually broke the concept down into two subcategories and then added them to make a total (A). We have simply recorded the rating in each of these subcategories accorded every test evaluated by CSE.[2] Since each CSE volume pertained to a particular age level of children, tests were often noted in more than one. This posed a problem since we wanted only one rating for each test. When multiple ratings did occur, we noted each rating and then averaged them for each category of test quality. When a test title happened to have subtests rated separately (and this occurred frequently), we chose the summary ratings for the title; if no summary existed, then we first noted each subtitle's rating in each of the six CSE volumes, and then averaged them all so as eventually to produce one rating score for each quality category for each title.[3]

The two subcategories of *Measurement Validity* are labeled by CSE as *Content and Construct* quality and *Concurrent and Predictive* quality. The first is a rating which CSE gives to each tests with respect to the percent of the test's goal actually assessed, the percentage of test items belonging in the test goal area, the empirical procedures for selecting test items, the theoretical support, the existence of divergent validity, factor analytic scores and the availability of experimental data on the test. These items are added by CSE and each test is given a rating which can range between 0 and 10, called *Content and Construct* quality.

Concurrent and Predictive quality is a rating based upon two characteristics: (1) the strength of the criterion correlation measures reported by the test authors or publishers, and (2) the strength of this validation correlation across time periods. From this subcategory, each test could receive a rating from 0 to 5. Thus, total (A) ratings of *Measurement Validity* contained a range of 0 to 15.

Normed Technical Excellence is broken down by CSE into six subcategories, and their total. The first three are evaluations of reliability: *Test-Retest Stability, Internal Consistency,* and *Alternate Form.* These are followed by subcategories called *Replicability, Range-Coverage,* and *Score Gradation.* The *Stability* subcategory reflects the level of correlation between test scores over time spans of one month or more. The consistency of items as measured by split-half, Kuder-Richardson or alpha coefficients make up the subcategory labeled *Internal Consistency.* The reliability of *Alternate Forms* of a test is evaluated by scoring the level of its

2. It might be mentioned that the CSE definition of validity did vary somewhat between the first and the more recent volume of test evaluations.

3. It is important to note that a test title was scored on both frequency and quality without reference to whether it was used on older or on younger children. This would have been a valuable addition, and when resources allow, it should be pursued.

alternate form correlation coefficients. If the test procedures are standardized in administration, scoring and interpretation, and if the characteristics are replicable of the groups upon which the standardization has been based, then the test would be rated high in *Replicability*. The *Range-Coverage* subcategory evaluates the adequacy of the ceiling and floor of the score distribution. The category of *Score Gradation* is a check on the capacity of a test to discriminate between groups, such as between centiles, grade equivalents or mental ages. The *Normed Technical Excellence* ratings are totaled to elicit a score ranging from 0 to 15. Finally, what we have done is to combine *Measurement Validity* and *Normed Technical Excellence* ratings to elicit a total score for each test with a range from 0 to 30.

TEST QUALITY AND ITS RELATIONSHIP TO TEST USAGE

Is there any relationship between the frequency with which a particular test or measure is cited by principal investigators in Federally-sponsored research proposals and the published ratings of the test's quality? A response to this question can be derived from the Spearman correlations displayed in Tables 4 and 5.

Table 4 illustrates the intercorrelations between each of the eleven indices of a test's quality. There is considerable variability, ranging from a low coefficient between the test's replicability and its stability ($r = .01$) to the high relationship between a test's range and its score gradation ($r = .85$ $p < .01$). As might be expected, each of the summary categories (A), (B) and (A+B) elicit very strong coefficients with the variables from which they are derived.

The bottom row in Table 4 contains the coefficients between the summary measure of test quality and frequency of use. By and large they are all positive, though of varying strengths and levels of statistical significance. The lowest is the

TABLE 3
Means, Standard Deviations, and Ranges of Quality for Tests In Selected Subject Areas

Test Category	Number of Titles	Quality Mean (Out of 30 Points Possible)	Standard Deviation	Quality Range (0-30 Possible)
Academic Achievement Batteries	19	14.2	3.7	7-19
Character and Personality	23	6.7	4.0	1-15
Non-Projective	43	7.6	3.1	2-14
Projective	5	3.0	1.7	2-6
Total	71	7.0	3.5	1-15
Spelling	2	10.5	5.0	7-14
Music	2	11.9	2.9	10-14

TABLE 3 (*Continued*)

Test Category	Number of Titles	Quality Mean (Out of 30 Points Possible)	Standard Deviation	Quality Range (0-30 Possible)
Foreign Languages	1	15.0	—	—
English	1	5.4	—	—
Spanish	1	6.8	—	—
Intelligence				
Group	16	10.5	3.9	3-19
Individual	19	10.3	4.1	5-19
Specific	6	9.7	2.6	6-13
Total	41	10.3	3.8	3-19
Mathematics	1	5.0	—	—
Algebra	1	12.0	—	—
Arithmetic	3	14.5	2.9	12-18
Industrial Arts	1	9.8	—	—
Listening Comprehension	1	5.0	—	—
Multi-Aptitude Batteries	4	13.2	3.2	9-16
Reading	11	12.6	4.4	2-17
Diagnostic	6	8.6	3.2	3-12
Miscellaneous	2	7.2	1.6	6-8
Oral	3	7.2	3.6	3-10
Readiness	19	9.3	4.2	3-16
Study Skills	5	9.9	1.3	8-11
Total	46	9.8	4.0	2-17
Biology	1	17.0	—	—
Sensory-Motor	7	7.9	2.7	3-11
Motor	1	3.8	—	—
Vocational Tests	2	9.5	4.3	6-13
Clerical	1	6.7	—	—
Interest Inventory	8	6.0	2.0	4.9
Manual Dexterity	2	4.5	.7	4-5
Mechanical Ability	2	9.0	7.8	4-15
Total	15	6.7	3.2	4-15
Speech and Hearing				
Hearing	1	6.0	—	—
Speech	9	6.4	3.3	3-12
Learning Disabilities	1	9.0	—	—

TABLE 4

Inter-Correlation Between Each Measure of Test Quality and Frequency of Test Use

	Content and Construct	Concurrent and Predictive	Total (A)	Stability	Internal Consistency	Alternate Form	Repli-cation	Range	Score Gradation	Total (B)	Total (A+B)	Frequency
C+C	—											
C+P	.39**	—										
Total (A)	.88**	.74**	—									
Stability	.21**	.38**	.33**	—								
Int. Consistency	.32**	.38**	.38**	.23**	—							
Alternate Form	.24**	.36**	.35**	.15	.24**	—						
Replication	.20*	.23**	.25**	.01	.32**	.26**	—					
Range	.27**	.41**	.38**	.20*	.49**	.30**	.41**	—				
Score Gradation	.30**	.35**	.36**	.21**	.45**	.24**	.35**	.85**	—			
Total (B)	.37**	.52**	.44**	.34**	.72**	.43**	.53**	.89**	.86**	—		
Total (A+B)	.67**	.70**	.79**	.39**	.67**	.46**	.47**	.78**	.75**	.88**	—	
Frequency	.11	.02	.09	.10	.17*	.22**	.05	.12	.18*	.17	.16*	—
$\bar{X}$	4.54	0.75	5.44	0.21	0.78	0.23	0.49	1.18	1.02	3.64	9.08	8.68
SD	1.60	0.97	2.11	0.59	1.06	0.59	0.46	1.08	0.82	3.00	4.29	20.56

*$p < .05$
**$p < .01$

TABLE 5

Correlations Between Each Measure of Test Quality and Frequency of Test Use, In Eight Subject Categories

Quality	Achievement Batteries (N = 19)	Vocational Skills (N = 15)	Total Intelligence Tests (N = 41)	Group Intelligence Tests (N = 16)	Individual Intelligence Tests (N = 19)	Personality (N = 71)	Reading (N = 46)	Sensory Motor (N =8)
Content and Construct	.29	.80*	−.04	−.18	.20	−.19	.13	.51
Concurrent and Predictive	.30	.64	−.07	−.11	.00	−.19	.05	.18
Total (A)	.48*	.21	.05	.14	.17	−.11	.13	.31
Stability	.08	a	.07	−.47	.30	−.21	−.40	.52
Internal Consistency	.41	.79*	−.02	.18	−.22	.02	.22	.58
Alternate Form	.37	a	−.09	.62	−.16	.28	.23	a
Replication	−.18	−.21	.41*	.65	.50*	.01	.01	.28
Range	−.10	.28	.42*	.13	.56*	−.15	−.05	.19
Score Gradation	.13	.28	.42*	.38	.38	−.01	.03	.01
Total (B)	.45	.11	.20	.06	.14	−.03	.09	.33
Total (A + B)	.49*	.13	.14	.23	.15	−.10	.13	.46

[a]Insufficient data to calculate a coefficient.
*p<.05.

relationship between frequency and concurrent and predictive quality ($r = .02$), but rather marked and significant relationships appear between frequency of use and a test's range, internal consistency, score gradation and two of the three summary measures ($r = .17$ and .16). Thus, taking all the 230 tests and measures mentioned in Federally-sponsored proposals on children and youth and evaluated by CSE, there is some relationship between a test's quality and its frequency of use.

However, when tests are broken down into categories of subject matter, differences emerge in the tendency to use high quality instruments. See Table 5. For example, among the 19 batteries of academic achievement, strong and statistically significant correlations can be found with two summary measures of quality: Total A ($r = .48\, p < .01$) and Total A+B ($r = .49\, p < .01$). This would indicate that in the case of achievement batteries, those rated the highest were generally mentioned more frequently in project proposals. But among tests of vocational skills, the impression is not so clear. Very strong and statistically significant relationships appear between frequency and content and construct quality, and also between frequency and internal consistency. But though consistently positive, none of the relationships with the three summary measures is statistically significant.

Relationships between the summary measures of quality and the frequency of intelligence test usage are also equivocal: all are positive but none statistically significant. By contrast, quite strong and significant coefficients appear between intelligence tests and the quality of their replication, range, and score gradation. At least researchers are not using the worst more frequently and, with respect to the criteria of quality used in this assessment, they are quite definitely using the better of those available.

In the case of sensory-motor skills, only eight of the test titles mentioned in Federal research proposals were evaluated. The coefficients are uniformly positive with many being quite strong. More titles would have to be correlated before the figures could be trustworthy.

The anomaly lies in the last two categories: tests of personality and tests of reading. Reading tests are quite numerous; 65 titles appeared in the proposals and one out of every ten projects included a test of reading. Nevertheless, despite the common interest in gauging reading skills, those tests selected for use appear to have little or no relationship to quality. However with respect to tests of personality, an additional query might be raised. Although no coefficient is statistically significant, in most categories of quality there is a *negative* relationship with frequency of use. If these data on the Federally-sponsored usage of reading and personality tests are representative of research in general, then at the very least they indicate (particularly in the case of personality tests) that the higher rated indices have no better chance of being utilized than those of poor quality.

SUMMARY AND IMPLICATIONS FOR POLICY

These data indicate three things. First, if taken as an undifferentiated unit, the better rated tests are generally used more frequently. But second, this

generalization is more true of some categories of tests than others. It is more true of achievement batteries, tests of vocational skills and tests of intelligence, in that order. Third, particular problems appear in the use of tests of reading and of personality. In these latter two categories, the higher-rated tests are not used more often.

It is evident that the normal process of selecting tests in which researchers choose the one they think is best, works more efficiently in some test categories than others. In some subjects there is more agreement on measurement definitions, more communication among researchers and consequently fewer titles. For example, this may account for the differences noted here between the field of academic achievement research and the research on personality.

The Interagency Panels may be able to serve a useful purpose with these data. Because the Panels house the collection of the most current and the most complete information on test usage, they could extend the function they now perform for ongoing research and research findings, to particular tests and measures. Now the Panels can answer a question from consumers in the field or from government agencies as to which tests are used more often. This in itself is unique and valuable. Furthermore, this test information can assist in efforts to encourage greater accumulation of research findings. As this paper has tried to demonstrate, the Panels can point out not only which tests are used more frequently, but which categories of tests might be especially deserving of attention in terms of quality.

CONTRIBUTORS

STEPHEN P. HEYNEMAN, Sociologist/Educator, Education Central Projects Staff, The World Bank, 1818 H Street, N.W., Washington, D.C. 20433.

PAMELA COPE MINTZ, Research Associate, Social Research Group, George Washington University, Washington, D.C. 20037.

REFERENCES

BUROS, O. K. *Tests in print, II.* Highland Park, N.J.: The Gryphon Press, 1974.

CHUN, K., COBB, S., & FRENCH, J. R., Jr. Measures for psychological assessment. Ann Arbor, Mich.: University of Michigan, 1975.

HERTZ, T. W., & MANN, A. J. *Toward interagency coordination: FY '75 federal research and development activities pertaining to early childhood.* Washington, D.C.: Social Research Group, The George Washington University, December, 1975.

HEYNEMAN, S. P. *Toward interagency coordination: FY '75 federal research and development activities pertaining to adolescence.* Washington, D. C.: Social Research Group, The George Washington University, December, 1975.

HOEPFNER, R. *CSE elementary school test evaluation.* Los Angeles, Calif.: UCLA Graduate School of Education, 1970.

HOEPFNER, R., CONNIFF, W. A., Jr., HUFANO, L., BASTONE, M., OGILVIE, V. N., HUNTER, R., & JOHNSON, B. L. *CSE secondary school test evaluations: grades 11 and 12.* Los Angeles, Calif.: UCLA Graduate School of Education, 1974(c).

HOEPFNER, R., CONNIFF, W. A., Jr., MCGUIRE, T. C., KLIBANOFF, L. S., STANGEL, G. F., LEE, H. B., & REST, S. *CSE secondary school test evaluations: grades 9 and 10.* Los Angeles, Calif.: UCLA Graduate School of Education, 1974(b).

HOEPFNER, R., CONNIFF, W. A., Jr., PETROSKO, J. M., WATKINS, J., ERLICH, O., TODARO, R. S., & HOYT, M. F. *CSE secondary school test evaluations: grades 7 and 8.* Los Angeles, Calif.: UCLA Graduate School of Education, 1974(a).

HOEPFNER, R., HEMENWAY, J., DEMUTH, J., TENOPYR, M. L., GRANVILLE, A. C., PETROSKO, J. M., KRAKOWER, J., SIBERSTEIN, R., & NADEAU, M. *CSE-RBS test evaluations: tests of higher-order cognitive, affective, and interpersonal skills.* Los Angeles, Calif.: UCLA Graduate School of Education, 1972.

HOEPFNER, R., STERN, C., & NUMMENDAL, S. G. *CSE-ECRC preschool/kindergarden test evaluations.* Los Angeles, Calif.: UCLA Graduate School of Education, 1971.

ROBINSON, J. P., & SHAVER, P. R. *Measures of social psychological attitudes.* Ann Arbor, Mich.: Survey Research Center, Institute for Social Research, August, 1969.