DOCUMENT RESUME

ED 138 644 TM 006 301

AUTHOR Newman, Dorothy C.; And Others
TITLE Teachers and Tests: A Concurrent Validity Study.
PUB DATE [Apr 77]
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, New York, April 5-7, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Rating; Achievement Tests; Correlation; Disadvantaged Youth; Elementary School Students; Grade Equivalent Scores; *Low Achievers; Primary Education; *Reading Achievement; Reading Level; *Reading Tests; *Standardized Tests; Student Evaluation; *Test Validity
IDENTIFIERS Iowa Tests of Basic Skills; Out of Level Testing

ABSTRACT

To determine the concurrent validity of a standardized test and its usefulness to educators, reading ratings from the Iowa Tests of Basic Skills were compared to teacher ratings and independently administered placement test ratings. Two hundred one primary children, the majority of whom read one or more years below grade level, took the ITBS by a grade testing plan. Correlation analysis, analysis of variance, and examination of difference scores supported the conclusion of low concurrent validity for the standardized test with low-achieving readers' instructional reading levels. Out-of-grade testing was recommended for low-achieving primary children. (Author)

Teachers and Tests: A Concurrent Validity Study

Dorothy C. Newman, John H. Neel, Patricia B. Campbell
Department of Educational Foundations
Georgia State University
Atlanta, Georgia

Paper presented at the annual meeting of the National
Council on Measurement in Education
New York City, April, 1977

Teachers and Tests: A Concurrent Validity Study

The usefulness of a standardized test depends on several factors such as the measure's validity, reliability, economy, and ease of interpretation (Stanley & Hopkins, 1972). To determine the concurrent validity of a standardized test and, hence, a necessary condition for its utility, researchers compared reading grade equivalent scores from a standardized test with teacher ratings and with placement test ratings. Interest focused on determining such validity for low-achieving primary grade readers in an inner city school when the children took the standardized test according to a graded testing plan. A graded testing plan calls for all children in a grade to take the same level of test regardless of the children's varying achievement levels.

After four years of classroom experiences and observations as a primary grade teacher, the principal author questioned the validity of standardized test scores with children's actual reading levels for these low-achieving readers tested under a graded testing policy. During classroom testing situations, observations showed that a child tended to guess answers or mark responses randomly when the standardized test was on or above his/her frustration level, or that level one or more years above the child's instructional reading level. The child who could not read the test experienced frustration, while the child who could read the test tended to take it as directed. Children who felt a need to guess answers or mark random responses were not being measured (Hieronymus & Lindquist, 1971). It appeared their responses were given not out of a knowledge base, but out of frustration. Consequently, their scores could not be interpreted as valid measures.

Normally, school systems use standardized test results in helping the teacher with tasks such as determining group and individual diagnoses and prescriptions, forming intraclass groupings, and assessing student growth. They use these results in helping administrators perform tasks such as assessing instructional programs, making decisions concerning planning and grouping, and comparing school units (Hieronymus & Lindquist, 1974). Therefore, determining whether or not a standardized test possesses concurrent validity with actual reading levels for the type of children identified is of the greatest import, since invalid scores are useless to teachers and administrators alike.

Researchers undertook the determination of the agreement or disagreement of standardized test reading ratings with two criterion rating sources: teacher judgments and reading placement test results. Previous studies had compared standardized test reading results with informal reading inventory results (Johns, 1972; McCracken, 1962; Sipay, 1964), so researchers felt the third rating source--teacher judgments--was necessary to verify the accuracy of results from a reading placement test administered independently by a researcher.

Looking at the three studies cited, it was difficult to draw definitive conclusions regarding standardized test validity for results were mixed. Also the students in these three studies were not classified as low-achieving readers. McCracken (1962) indicated that 63% of children in his study would have been grouped on frustration reading levels if the standardized test scores alone had been the basis for group mem cisions. Sipay (1964) discovered that standardized test scores tend o overestimate instructional reading levels, a position commonly held among educators. At the same time they underestimated frustration reading levels.

Johns (1972) showed that 16% of the children in his study were rated one grade level or more above their instructional reading levels by the standardized test.

Instruments

The eight teachers taking part in the study administered subtests Vocabulary and Reading Comprehension of the Iowa Tests of Basic Skills (1971) for the Primary Battery, Level 8, Form 5 and the Regular Battery, Level 9, Form 5 as part of the yearly testing program in a Southern urban school system. A researcher administered the Macmillan Reader Placement Test (1972) individually to children. This test consisted of two parts: vocabulary recognition and reading selection comprehension.

Methods and Results

Two questions were posed for study: Do teacher ratings and placement test ratings agree on children's reading levels? Do standardized test ratings possess concurrent validity for the two criterion source ratings of children's reading levels? To answer these questions, researchers proposed the statistical null hypotheses that there would be no significant differences between reading ratings from (1) teacher judgments and placement test results, (2) teacher judgments and standardized subtest vocabulary results, (3) teacher judgments and standardized subtest comprehension results, (4) placement test results and standardized subtest vocabulary results, and (5) placement test results and standardized subtest comprehension results.

A sample of 201 second and third grade children in an urban elementary school was chosen for several reasons: the majority of children read one year or more below grade level according to school records; the children were administered a standardized test according to a graded testing plan; and the children, teachers, and test scores were made available

to researchers by the school system for the validation study. The school chosen was a Title I school located in a low income housing project. The sample included only 201 of 230 children enrolled because of the loss of children who did not take all tests used in the comparisons. Of the children in the sample, 76% read on a level half a year or more below grade level, while 59% read on a level one year or more below grade level according to school records.

Primary children were selected as the target group because of a need for an examination of standardized test concurrent validity for this age group. Focusing on reading ratings seemed to be highly appropriate for children in the primary grades where great emphasis normally is placed on reading instruction. Two major advantages also influenced the limiting of the study to an examination of reading ratings. Previous classroom experiences in administering reading placement tests was an advantage to researchers, and the fact that teachers already had judged their students' reading levels in the course of regular instruction was an advantage to them.

The Iowa Tests of Basic Skills (ITBS) were administered by the eight classroom teachers to all children, except the mentally and physically handicapped and those habitually absent, during May, 1974. In the weeks immediately preceding and following the ITBS administration, a researcher administered the Macmillan Reader Placement Test. The placement test was based on the children's basal reader series and was administered individually. Placement test administration occurred independently of teacher ratings.

After the placement tests were completed, each of the eight teachers was given a list of her students on which to indicate reading grade levels.

Teachers were asked to make this decision for each child according to the child's basal reader instructional level. For example, a child reading in a second grade first semester reader would be rated 2^1 as his/her reading level. To insure comparability of scores, the grade level ratings from all three sources were classified as seen in Table 1. Standardized test ratings are grade equivalent scores. This classification allowed each child to vary within one-half school year, or within five school months.

TABLE 1

Classification of Grade Level Ratings

Teacher Ratings	Placement Test Ratings	Standardized Test Ratings	Study Classification
R	R	0.0-0.9	.5
Pp/P	Pp/P	1.0-1.4	1.0
1	1	1.5-1.9	1.5
2^1	2^1	2.0-2.4	2.0
2^2	2^2	2.5-2.9	2.5
3^1	3^1	3.0-3.4	3.0
3^2	3^2	3.5-3.9	3.5
4	4	4.0-4.9	4.0
5	5	5.0-5.9	5.0
6	6	6.0-6.9	6.0

Note.--Fourth, fifth, and sixth grade readers cover one school year. Children placed in these readers may vary within this range.

The sample was divided into eight homeroom units, intact groups extant at the school, for data analysis. Since teacher and placement test ratings included vocabulary and comprehension skills, an examination of both vocabulary and comprehension subtests of the ITBS was needed.

As a measure of the concurrent validity between the standardized test ratings and the criterion source ratings, correlation coefficients were obtained for grade level ratings between every pair of the three rating source combinations as shown in Table 2 (see Appendix A for correlations by classes).

TABLE 2

Correlations Between Rating Sources: Total Sample

	Vocabulary	Comprehension	Teacher	Placement Test
Vocabulary	1.00			
Comprehension	.40	1.00		
Teacher	.51	.38	1.00	
Placement Test	.55	.35	.91	1.00

n = 201.

Stanley and Hopkins (1972) explained concurrent validity as the extent of correlation between two concurrently obtained criteria. In this sense, the ITBS ratings possessed concurrent validity only to the extent which they correlated with teacher and placement test ratings. Correlations were strongest for comparisons of teacher ratings with placement test ratings and were weakest for comparisons of comprehension subtest ratings with both teacher and placement test ratings.

Grade level scores determined by the three rating sources were also studied for each child. Several interesting facts were revealed. Teacher and placement test ratings differed one year or more in only 5% of the cases. ITBS vocabulary ratings overestimated teacher ratings one year or more in 15% of the cases and underestimated teacher ratings in 20% of the cases. ITBS comprehension ratings overestimated teacher ratings one year or

more in 15% of the cases and underestimated teacher ratings one year or more in 25% of the cases. ITBS vocabulary ratings overestimated placement test ratings one year or more in 12% of the cases and underestimated placement test ratings one year or more in 20% of the cases. ITBS comprehension ratings overestimated placement test ratings one year or more in 15% of the cases and underestimated placement test ratings one year or more in 25% of the ratings.

Such an examination of difference scores between rating sources showed that standardized test scores neither overestimated nor underestimated teacher and placement test ratings with any consistency. Standardized test ratings differed from the two criterion source ratings one year or more in either direction in 32% to 40% of the cases, depending on the comparison.

An analysis of variance using a randomized block design with the rating sources as treatments and the four classes of each grade as blocks, was conducted for each grade at the .05 significance level. The analysis of variance in grade two resulted in rejection of the null hypothesis for comparisons between (1) teacher and ITBS vocabulary ratings, (2) teacher and ITBS comprehension ratings, and (3) placement test and ITBS comprehension ratings. In grade three the null hypothesis was not rejected for comparisons between every pair of raters. Results for these tests are given in Appendix B.

Conclusions

One of the most interesting findings of the study was the fact that for the group of children tested standardized test ratings did not consistently overestimate instructional reading levels, but instead the standardized test ratings both overestimated and underestimated

instructional reading levels as determined by teacher judgments and placement test ratings.

Differences between the standardized test ratings and the two criterion source ratings of one year or more in 32% to 40% of the cases raised questions as to the standardized test's usefulness for teachers and other professional educators. The fact that the differences were not consistently in one direction further clouded the issue of their usefulness, since a consistent difference in either direction could provide information that could be used.

Correlation analysis also supported the notion of little concurrent validity for the standardized test ratings with instructional reading levels for the low-achieving readers in the study. A test's concurrent validity may be measured by the extent to which it correlates with a concurrently obtained criterion. In the present study, teacher and placement test ratings correlated to a substantial degree for all eight classes and the total sample with correlations from .85 to .96. Ratings from teachers and the ITBS vocabulary subtest correlated from .35 to .68 for the classes with a correlation of .51 for the entire sample.

Teacher and ITBS comprehension ratings correlated from .13 to .64 with a total sample correlation of .38. The overall correlation probably was misleading since classes one through four, the second grade classes, had correlations of .58 to .64 while classes five through eight, the third grade classes, had correlations of .13 to .37. This same trend was evident in the correlation between placement test and ITBS comprehension ratings. Correlations between ratings from these two sources for the classes ranged from .12 to .61 with a total sample correlation of .35.

Speculation about the difference between the second and third grade correlations for the ITBS comprehension ratings and the two criterion source ratings led to a closer examination of these ratings. It revealed that more second graders read on a second grade level, 56%, than third graders read on a third grade level, 29%. Lower correlations were obtained between ITBS comprehension ratings and both teacher and placement test ratings in the third grade where a smaller percentage of children read on grade level. This phenomenon is consistent with the original classroom observation that children who could not read a standardized test obtained scores markedly varied from teachers' estimates of instructional reading levels.

For correlations between placement test and ITBS vocabulary ratings for the eight classes the lower correlations again occurred in two third grade classes, although the trend noted above was not as clearly delineated for the ITBS vocabulary subtest as for the ITBS comprehension subtest comparisons. For the total sample, placement test and ITBS vocabulary ratings correlated .55.

Standardized test ratings correlated with both teacher ratings and placement test ratings within a range of from .12 to.68. None of these correlations, however, approached the strength of the correlations between teacher and placement test ratings, which ranged from .85 to .96. The obtained correlations supported the hypotheses of agreement between teacher and placement test ratings and disagreement between the two criterion source ratings and the standardized test ratings.

Hypothesis testing was not a particularly fruitful technique in the study. However, the research questions more appropriately concerned information about individuals rather than group means only.

Educational Importance

The administration of standardized tests constitutes a major portion of many school system testing programs. They are designed to help teachers and administrators in their professional tasks. If obtained scores are not valid, however, the information they contain is useless to teachers and administrators for instructional planning and implementation. At worst, invalid scores can be used in a manner harmful to a child if the user of the scores assumes them to be valid. Results of the present study called into question the use of one standardized test for young, low-achieving readers tested according to a graded testing policy administered by the school system. Particularly questionable were the results of the comprehension subtest for third graders in the sample.

It would be more appropriate to administer the ITBS in accordance with either of two alternative plans suggested by the test publishers (Hieronymus & Lindquist, 1974). An out-of-grade plan calls for administration of one test level to all children in a grade or subgrouping within a grade with the test level being either lower or higher than actual grade placement, whichever is more suitable. An individualized plan calls for the administration of an appropriate test level for each child. Implementing either of the two plans precludes grade level comparisons with a norming group; however, administration of inappropriate test levels seems to yield questionable results. It is suggested that out-of-grade norms be developed for local school systems to gain more useful information on an immediate level regarding pupil achievement and instructional programming. Because of time and cost factors, an out-of-grade plan probably is more feasible.

Until alternatives to the graded testing plan are found for low achievers who are classed by grade levels according to age, standardized test score interpretations should be done with a certain amount of caution. It is important that professio[illegible]ealize the pointlessness of giving standardized test[illegible] cannot read them. S[illegible]ch testing experiences can only resu[illegible]frustration for children and useless information for teachers.

References

Hieronymus, A. N., & Lindquist, E. F. Teacher's guide for administration, interpretation, and use: Iowa Tests of Basic Skills, regular battery, forms 5 & 6. Boston: Houghton Mifflin, 1971.

Hieronymus, A. N., & Lindquist, E. F. Manual for administrators, supervisors, and counselors: Iowa Tests of Basic Skills. Boston: Houghton Mifflin, 1974.

Johns, J. L. Do standardized tests rate pupils above their instructional reading levels? Reading Teacher, 1972, 25(6), 569.

McCracken, R. A. Standardized reading tests and informal reading inventories. Education, 1962, 82, 366-369.

Sipay, E. R. A comparison of standardized reading scores and functional reading levels. Reading Teacher, 1964, 17, 265-268.

Sipay, E. R. The Macmillan Reader Placement Test, forms 1 and 2. New York: Macmillan, 1972.

Stanley, J. C., & Hopkins, K. D. Educational and psychological measurement and evaluation. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

APPENDIX A

Correlation Tables by Classes

TABLE A

Correlations Between Rating Sources: Class 1

	Vocabulary	Comprehension	Teacher	Placement Test
Vocabulary	1.00			
Comprehension	.57	1.00		
Teacher	.45	.58	1.00	
Placement Test	.53	.56	.91	1.00

n = 25

TABLE B

Correlations Between Rating Sources: Class 2

	Vocabulary	Comprehension	Teacher	Placement Test
Vocabulary	1.00			
Comprehension	.[illegible]	.[illegible]		
Teacher	.45	.61	1.00	
Placement Test	.50	.57	.85	

n = 23

TABLE C

Correlations Between Rating Sources: Class 3

	Vocabulary	Comprehension	Teacher	Placement Test
Vocabulary	.00			
Comprehension	.56	1.00		
Teacher	.68	.64	1.00	
Placement Test	.66	.56	.87	1.00

n = 24

TABLE D

Correlations Between Rating Sources: Class 4

	Vocabulary	Comprehension	Teacher	Placement Test
Vocabulary	1.00			
Comprehension	.58	1.00		
Teacher	.67	.63	1.00	
Placement Test	.68	.61	.95	1.00

n = 23

TABLE E

Correlations Between Rating Sources: Class 5

	Vocabulary	Comprehension	Teacher	Placement
Vocabulary	1.			
Comprehension	.4-	1.00		
Teacher	.5-	.34	1.00	
Placement Test	.57	.33	.91	1.00

n = 25

TABLE F

Correlati[illegible]tween Rating Sources: Class 6

	Vocabu	Comprehension	Teacher	Placement Test
Vocabulary	1.			
Comprehension	..	1.00		
Teacher	.43	.37	1.00	
Placement Test	.36	.31	.96	1.00

n = 26

TABLE G

Correlations Between Rating Sources: Class 7

	Vocabulary	Comprehension	Teacher	Placement Test
Vocabulary	1.00			
Comprehension	.27	1.00		
Teacher	.35	.24	1.00	
Placement Test	.35	.12	.91	1.00

n = 29

TABLE H

Correlations Between Rating Sources: Class 8

	Vocabulary	Comprehension	Teacher	Placement Test
Vocabulary	1.00			
Comprehension	-.08	1.00		
Teacher	.61	.13	1.00	
Placement Test	.64	.19	.91	1.00

n = 26

APPENDIX B

Analysis of Variance Tables by Grades

TABLE A
Analysis of Variance Table
Teacher and Placement Test: Grade Two

Source	SS	df	MS	F
Between Raters	0.0009	1	0.0009	0.0018
Between Classes	5.3300	3	1.7770	3.4721**
Residual	94.6791	185	0.5118	
Total	100.0100	189		

**p < .01

TABLE B
Analysis of Variance Table
Teacher and Vocabulary Subtest: Grade Two

Source	SS	df	MS	F
Between Raters	2.10	1	2.1000	4.0268*
Between Classes	4.67	3	1.5600	2.9914**
Residual	96.47	185	0.5215	
Total	103.24	189		

*p < .05
**p < .01

TABLE C
Analysis of Variance Table
Teacher and Comprehension Subtest: Grade Two

Source	SS	df	MS	F
Between Raters	6.16	1	6.1600	50.0813**
Between Classes	11.81	3	3.9400	32.0325**
Residual	22.76	185	0.1230	
Total	40.73	189		

**p < .01

TABLE D
Analysis of Variance Table
Placement Test and Vocabulary Subtest: Grade Two

Source	SS	df	MS	F
Between Raters	2.10	1	2.1000	3.
Between Classes	2.28	3	0.7600	1.2230
Residual	114.90	185	0.6211	
Total	119.28	189		

TABLE E
Analysis of Variance Table
Placement Test and Comprehension Subtest: Grade Two

Source	SS	df	MS	F
Between Raters	6.16	1	6.1600	29.3890**
Between Classes	11.70	3	3.9000	18.6069**
Residual	38.77	185	0.2096	
Total	56.63	189		

**p < .01

TABLE F
Analysis of Variance Table
Teacher and Placement Test: Grade Three

Source	SS	df	MS	F
Between Raters	0.14	1	0.1400	0.2642
Between Classes	55.65	3	18.5500	35.0066**
Residual	109.69	207	0.5299	
Total	165.48	211		

**p < .01

TABLE G
Analysis of Variance Table
Teacher and Vocabulary Subtest: Grade Three

Source			MS	F
Between Raters	0.02	1	0.0200	0.0309
Between Classes	1.39	3	0.4600	0.7101
Residual	134.09	207	0.6478	
Total	135.50	211		

TABLE H
Analysis of Variance Table
Teacher and Comprehension Subtest: Grade Three

Source	SS	df	MS	F
Between Raters	0.48	1	0.4800	0.7157
Between Classes	6.01	3	2.0000	3.1070*
Residual	133.25	207	0.6437	
Total	139.74	211		

*p $\leq$ .05

TABLE I
Analysis of Variance Table
Placement Test and Vocabulary Subtest: Grade Three

Source	SS	df	MS	F
Between Raters	0.27	1	0.2700	0.4189
Between Classes	1.74	3	0.5800	0.8998
Residual	133.43	207	0.6446	
Total	135.44	211		

TABLE J
Analysis of Variance Table
Placement Test and Comprehension Subtest: Grade Three

Source	SS	df	MS	F
Between Raters	0.08	1	0.0800	0.1251
Between Classes	6.50	3	2.1700	3.3927*
Residual	132.39	207	0.6396	
Total	138.97	211		

*$p < .05$