ED 106 352 TM 004 468

AUTHOR	Roudabush, Glenn E.
TITLE	Estimating Normative Scores from a Criterion-Referenced Test.
PUB DATE	[Apr 75]
NOTE	18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)
EDRS PRICE	MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS	Achievement Tests; Correlation; *Criterion Referenced Tests; Equated Scores; Measurement Techniques; Multiple Regression Analysis; *Norm Referenced Tests; Prediction; Raw Scores; *Reading Tests; *Scores; Scoring Formulas; *Standardized Tests; Test Reliability; Test Validity; Weighted Scores
IDENTIFIERS	California Achievement Test; Prescriptive Reading Inventory

ABSTRACT

The objective of this study was to show that standardized reading scores could be adequately estimated from scores on a criterion-referenced test in reading. This would reduce classroom test time, while, at the same time, provide the kinds of information teachers need to guide instruction, and the kinds of information administrators require for making decisions regarding education programs. Stepwise regression and equipercentile equating were used to estimate scores from the criterion-referenced scores. The results show that it is possible to estimate normative scores from a broad based criterion-referenced test in reading. (Author)

ED106352

ESTIMATING NORMATIVE SCORES FROM A CRITERION-REFERENCED TEST[1]

Glenn E. Roudabush

CTB/McGraw-Hill

TM 004 468

A paper presented at the

American Educational Research Association

meetings in Washington, D.C., April 1, 1975

In the literature on educational measurement discussions frequently appear on the differences between standardized, norm-referenced tests and criterion-referenced tests. The differences are in their construction, interpretation, and use. Norm-referenced achievement tests are constructed to measure broad educational goals, and items are selected to discriminate the amount of knowledge or skill a student has in a particular achievement domain. Construction procedures tend to spread kids out on a continuum and bring out individual differences where they exist (with respect to performance on the particular test). Criterion-referenced tests are constructed to measure specific educational objectives, and items are selected to discriminate between individuals who have or have not mastered these particular objectives. Construction procedures tend to maximize the instructional effects on scores rather than the individual differences of students. Norm-referenced tests are useful for long-term evaluation of educational progress, while criterion-referenced tests are useful for evaluating short-term instruction and, therefore, for assisting teachers in diagnosing strengths and weaknesses of students and planning their instruction.

Criterion-referenced test information is most useful to the classroom teacher in giving insight into, and guidance for, instruction. Such tests provide diagnostic and prescriptive information about each student that allows the teacher to plan instruction for groups and individual students best suited to meet their individual instructional needs. There is, however, a continuing demand and need by educational administrators, legislatures, and the general public for comparative or normative data on students in order to make intelligent decisions about the allocation of resources and in order to know how they stand with respect to national, state, or district performance.

Time taken from instructional time to administer standardized, norm-referenced tests in the classroom, which have no direct effect on instruction, is perceived by teachers and students as wasted time. If it is possible to use the same instrument to provide teachers the kind of information they need, that is, criterion-referenced information, and, at the same time provide administrators the kind of information they need, that is, norm-referenced information, then the time and effort put into testing by teachers and students alike will be perceived as useful and less threatening to both.

It is possible, of course, to norm a criterion-referenced test or to criterion-reference a norm-referenced test thus using the same test for both purposes. The consensus, however, seems to be that the differences between the two kinds of tests are such that a criterion-referenced norm-referenced test will be a poor substitute for a well constructed criterion-referenced test and that a normed criterion-referenced test will be a poor substitute for a well constructed norm-referenced test (see Hambleton & Novick, 1973; Messick, 1974).

We have been conducting studies to determine the relationships between the two types of tests and have found that by using regression analyses and equating techniques, a good, comprehensive criterion-referenced test can produce normative test results about as well as a norm-referenced test. It is interesting to note that our data show that the reverse is not true.

METHOD

The tests used in this study were the Reading Vocabulary, Reading Comprehension, and Reading Total scores from the <u>California Achievement</u>

Tests, 1970 Edition (CAT-70), a well-known nationally normed achievement series, and the Prescriptive Reading Inventory (PRI), a comprehensive criterion-referenced test of reading skills measuring about ninety objectives in four overlapping levels covering most of what is taught in reading from Grades 1.5 through 6. Both tests are published by CTB/McGraw-Hill.

The data for this study were collected in the fall of 1972 as part of a larger overall study of the PRI. The data collected were as follows:

Grade	Level	CAT-70 Level	Ethnic Code	No. of Cases
1.5	A	1	Stand	555
* 2.2	A	1	Stand	963
2.2	B	1	Stand	685
2.2	B	1	Black	935
* 3.2	B	1	Black	916
* 3.2	B	2	Stand	742
3.2	C	2	Stand	615
* 4.2	C	2	Stand	993
4.2	D	2	Stand	539
* 5.2	D	3	Stand	1498
6.2	D	3	Stand	1773

The procedure was to select one grade/level combination for each level of the PRI plus an additional data set from the black sample at level B of the PRI for the regression analysis. The selected grade/level combinations are indicated by asterisks in the table above. For each of these cells (grade/level combinations), 70% of the available data was randomly selected as the regression sample and the remaining (random) 30% was used for cross validation. Using a stepwise regression program, weights for prediciting the three CAT-70 raw scores from PRI objectives scores were obtained. The weights were cross validated with the remaining 30% of the data in each of these cells and, as a more stringent test of validity, the same weights were validated using a random 30% of the data from adjacent grades where the same level of the PRI and CAT-70 had been administered. There were, then, five regression analyses and nine cross validations. Some additional analyses were done using CAT-70 standard scale scores, but these analyses will be described below.

RESULTS

The results of the raw score regression analyses and cross validations thus far described are summarized in Table 1. In reading this table, note that the correlations under the regression analysis column are multiple correlations from the regression analyses. The correlations under the cross validation column are correlations between predicted and obtained CAT-70 reading scores in the validation samples, and the correlations under the alternate form CAT correlations column are the simple correlations between the reading scores from Form A and Form B of CAT-70. This table speaks pretty well for itself. The validity coefficients are quite high,

Insert Table 1 about here

sometimes exceeding both the multiple correlations from the regression analyses and the alternate form correlations. They are somewhat lower for the black sample, particularly when computed on data from a different grade. This may be a consequence of lower reliabilities for the Grade 2 black scores. There are also some marginally large differences between the actual CAT-70 means and the means of the predicted CAT scores when the weights are applied across grades. This occurs for the black sample and at Level D. The largest difference is -3.45 raw score points for Reading Total in the black sample. This difference represents about 7 percentile points or 1 to 2 months in grade equivalent score. The differences in means for cross validation at the same grade level are all less than one raw score point. Overall, these data suggest that the predicted CAT-70 reading scores from the PRI are about as good as an alternate form of CAT-70 itself.

Scale Score Analysis

Having proved to ourselves that predicting normative reading scores from the PRI was quite feasible and practical, we now wanted to obtain a single regression equation for each of the four levels of the PRI that would optimally predict CAT-70 scale scores, which are independent of the CAT level and from which derived scores (percentiles, grade equivalents) are easily obtainable. We also wanted to investigate further the equatability and scalability of the predicted scale scores.

We first converted the CAT-70 raw scores to scale scores, then pooled all of the data for a given level of the PRI to rerun the regression analyses. For Level A this included first and second grade data, for Level B second and third grade data for both the standard and black samples, for Level C third and fourth grade data, and Level D fourth, fifth, and sixth grade data. The four regression analyses were run and the weights obtained were adjusted to give the same mean and standard deviation for the actual and predicted scale scores and these were then applied in turn to each of the groups making up the data pool for that level of the PRI. The results of these analyses are shown in Table 2. Note first that the correlations hold up nicely, as would be expected. There are some differences in actual means and the means of the predictions. These, however, are not serious. The largest difference is -6.3 scale score units which occurs in

Insert Table 2 about here

Reading Comprehension for the Grade 2 standard sample. This difference represents slightly more than one raw score unit which is 1 or 3 percentile points or about one month in grade equivalent score.

In addition to the regression analyses, we obtained distributions of the actual and predicted scale scores for each group and of the differences between them (actual minus predicted). These distributions show some interesting properties of the predicted scale scores. The distributions of obtained scale scores can have data only at particular points on the scale score continuum corresponding to particular raw scores. These points may be separated by two or three scale score points near the middle of the distribution or by twenty or more scale score points near the ends of the distribution; that is, for obtained scale scores, missing an item (or getting an additional item correct) may change the scale score obtained by as much as twenty or more points. The predicted scale scores are based on a weighted composite of 30 to 35 objective scores each made up of three to five items.

For this reason, any scale score (including fractional ones) within the range of the test are possible. A change in performance on one or several items in the PRI will not change the predicted scale score very much. Typically in these distributions, there are data at every scale score point throughout the range of the test except at the ends of the distributions where the frequencies fall to zero abruptly. This effect may be due to the fact that to obtain a very high scale score a student must pass a large number of items, rather than just getting one two more items correct than the other students in the sample and, similarly, in order to get a very low predicted scale score a student must fail a large number of items. Assuming that the test is reasonably within the functional range for the students in the sample, either of these events is unlikely. Figure 1 shows this effect very nicely and is typical of all of the group distributions. In this

---

Insert Figure 1 about here

---

figure, the obtained and predicted scale scores are plotted against the normal deviates from the distributions for one test and one grade. The over prediction at the low end and the under prediction at the high end of the distribution are clear.

The obtained scale score distribution forms practically a straight line and these scores are approximately normally distributed. The distribution of predicted scale scores is platikurtic and throughout most of the range of the test is more like a uniform distribution than the normal. The predicted scale scores rank order students very well--better than do the obtained scale scores.

Summary statistics from the distributions of difference scores are shown in Table 3. In looking at these statistics, recall that these differences are between fixed points on the scale score continuum representing corresponding raw score points and scores which range through all scale score points within the range of the test. Also bear in mind the over and

Insert Table 3 about here

under predictions at the low and high ends of the distributions, respectively. Though most of the mean differences fall respectably close to zero, there is considerable variation in the accuracy of predicting individual scores. The standard deviations range from about 20 to 44 scale score points. Before we at CTB attempt to make any predictions of individual scores, we will attempt to improve accuracy by doing an equipercentile equating of the distributions of obtained and predicted scale scores.

References

California achievement tests, 1970 edition. Monterey, CA: CTB/McGraw-Hill, 1970.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and method for criterion-reference tests. Journal of Educational Measurement, 1973, 10, 159-170.

Messick, S. The standard problem: meaning and values in measurement and evaluation. RB 74-44. Princeton, NJ: Educational Testing Service, 1974.

Prescriptive reading inventory. Monterey, CA: CTB/McGraw-Hill, 1972.

Footnote

1. I wish to express my appreciation to Merrill E. Guest for his assistance in preparing this paper. In particular, he prepared Figure 1 which was most helpful in interpreting the distributional data.

TABLE 1. Results of the Raw Score Regression Analysis and Cross Validation

Data		Regression Analysis			Cross Validation			Alternate Form CAT Correlations		
		Vocab	Comp	Total	Vocab	Comp	Total	Vocab	Comp	Total
Grade 2, CAT 1, PRI A	- Correlations	.856	.846	.893	.829	.854	.906	.860	.770	.855
	- Stand. error	6.490	3.384	8.011						
	- N	657	653	653	268	264	264			
	- Act. mean				76.55	12.61	89.28			
	- Pred. mean				76.15	12.81	89.13			
	- Difference				.40	-.20	.15			
Grade 1, CAT 1, PRI A	- Correlations				.798	.724	.840			
(Weights from	- N				130	131	129			
grade 2 - lagged	- Act. mean				76.29	11.86	88.26			
back)	- Pred. mean				75.83	12.29	88.30			
	- Difference				.46	-.43	-.04			
Grade 3, CAT 1, PRI B	- Correlations	.768	.810	.830	.773	.828	.833			
(Black sample)	- Stand. error	9.532	3.156	10.891						
	- N	674	667	667	214	212	212			
	- Act. mean				75.00	13.51	88.47			
	- Pred. mean				74.50	13.52	88.15			
	- Difference				.50	-.01	.32			

TABLE 1. (Continued)

Data		Regression Analysis			Cross Validation			Alternate Form CAT Correlations		
		Vocab	Comp	Total	Vocab	Comp	Total	Vocab	Comp	Total
Grade 2, CAT 1, PRI B	- Correlations				.632	.679	.701			
(Black sample -	- N				199	198	195			
weights from	- Act. mean				64.12	9.83	74.45			
grade 3 - lagged	- Pred. mean				66.92	10.61	77.90			
back)	- Difference				-2.80	-.78	-3.45			
Grade 3, CAT 2, PRI B	- Correlations	.859	.844	.891	.729	.800	.855	.828	.787	.858
	- Stand. error	3.525	5.644	7.402						
	- N	589	589	589	227	228	230			
	- Act. mean				33.70	31.22	64.93			
	- Pred. mean				33.85	31.56	65.51			
	- Difference				-.15	-.34	-.58			
Grade 4, CAT 2, PRI C	- Correlations	.750	.841	.865	.692	.837	.855	.816	.790	.858
	- Stand. error	4.064	5.329	7.384						
	- N	666	666	666	215	209	204			
	- Act. mean				[illegible].35	35.07	70.62			
	- Pred. mean				35.44	34.18	70.05			
	- Difference				-.09	.89	.57			

TABLE 1. (Continued)

Data		Regression Analysis			Cross Validation			Alternate Form CAT Correlations		
		Vocab	Comp	Total	Vocab	Comp	Total	Vocab	Comp	Total
Grade 3, CAT 2, PRI C	- Correlations				.805	.823	.867			
(Weights from	- N				154	150	147			
grade 4 - lagged	- Act. mean				33.36	30.59	63.98			
back)	- Pred. mean				33.79	30.91	64.80			
	- Difference				-.43	-.32	-.82			
Grade 5, CAT 3, PRI D	- Correlations	.868	.850	.901	.890	.824	.900	.848	.837	.895
	- Stand. error	4.397	4.306	6.990						
	- N	1082	1082	1081	413	413	414			
	- Act. mean				25.25	24.51	49.79			
	- Pred. Mean				25.30	24.78	50.09			
	- Difference				-.05	-.27	-.30			
Grade 6, CAT 3, PRI D	- Correlations				.893	.847	.909			
(Weights from	- N				551	547	546			
grade 5 - lagged	- Act. mean				28.98	27.71	56.69			
forward)	- Pred. mean				27.51	26.74	54.26			
	- Difference				1.47	.97	2.43			

TABLE 2. Results of the Regression Analysis and Sub-Group Validation Using Scale Scores and Adjusted Weights

	CAT-70 Vocab.	CAT-70 Comp.	CAT-70 Total
PRI Level A			
Alternate form correlations	.860	.770	.855
Multiple correlations	.841	.792	.858
Standard Error of Estimate	21.0	32.7	22.0
Number of cases	1415	1411	1407
Grade 1.5 correlations	.823	.733	.834
Grade 1.5 N's	412	429	412
Grade 1.5 actual means	328.0	325.1	313.2
Grade 1.5 mean of predictions	327.0	325.8	313.0
Grade 1.5 difference (act. - pred.)	1.0	-.7	.2
Grade 2.2 correlations	.854	.824	.877
Grade 2.2 N's	852	865	844
Grade 2.2 actual means	332.7	335.3	320.3
Grade 2.2 mean of predictions	333.0	336.5	320.5
Grade 2.2 difference (act. - pred.)	-.3	-.8	-.2
PRI Level B			
Alternate form correlations	.828	.787	.858
Multiple correlations	.846	.801	.862
Standard Error of Estimate	27.1	37.5	29.3
Number of cases	3308	3297	3292
Grade 2.2 standard sample correlations	.846	.792	.858
Grade 2.2 standard sample N's	640	642	639
Grade 2.2 standard sample actual means	340.4	345.2	329.3
Grade 2.2 standard sample mean of pred.	338.2	351.5	329.2
Grade 2.2 standard sample diff.	2.2	-6.3	.1
Grade 2.2 black sample correlations	.708	.619	.719
Grade 2.2 black sample N's	707	708	704
Grade 2.2 black sample actual means	296.2	306.1	281.3
Grade 2.2 black sample mean of pred.	296.5	303.5	283.1
Grade 2.2 black sample diff.	-.3	2.6	-1.8
Grade 3.2 standard sample correlations	.774	.781	.817
Grade 3.2 standard sample N's	801	800	799
Grade 3.2 standard sample actual means	370.1	393.8	369.2
Grade 3.2 standard sample mean of pred.	373.1	396.1	369.7
Grade 3.2 standard sample diff.	-3.0	-2.3	-.5
Grade 3.2 black sample correlations	.814	.774	.838
Grade 3.2 black sample N's	846	826	831
Grade 3.2 black sample actual means	328.2	342.8	318.1
Grade 3.2 black sample mean of pred.	325.5	339.3	315.5
Grade 3.2 black sample diff.	2.7	3.5	2.6

TABLE 2. (Continued)

	CAT-70 Vocab.	CAT-70 Comp.	CAT-70 Total
PRI Level C			
Alternate form correlations	.816	.790	.858
Multiple correlations	.802	.821	.861
Standard Error of Estimate	30.1	36.8	30.2
Number of cases	1566	1563	1562
Grade 3.2 correlations	.805	.787	.845
Grade 3.2 N's	505	515	520
Grade 3.2 actual means	369.3	393.8	367.9
Grade 3.2 mean of predictions	368.9	391.4	367.1
Grade 3.2 difference (act. - pred.)	.4	2.4	.8
Grade 4.2 correlations	.782	.818	.853
Grade 4.2 N's	754	759	764
Grade 4.2 actual means	395.9	427.4	401.3
Grade 4.2 mean of predictions	396.8	427.9	401.1
Grade 4.2 difference (act. - pred.)	-.9	-.5	.2
PRI Level D			
Alternate form correlations	.848	.837	.895
Multiple correlations	.855	.834	.882
Standard Error of Estimate	34.4	39.6	33.1
Number of cases	3799	3796	3794
Grade 4.2 correlations	.698	.741	.786
Grade 4.2 N's	590	587	587
Grade 4.2 actual means	399.9	430.0	406.6
Grade 4.2 mean of predictions	402.4	428.3	405.9
Grade 4.2 difference (act. - pred.)	-2.5	1.7	.7
Grade 5.2 correlations	.860	.834	.886
Grade 5.2 N's	1389	1378	1377
Grade 5.2 actual means	426.7	452.5	429.7
Grade 5.2 mean of predictions	429.0	453.9	432.4
Grade 5.2 difference (act. - pred.)	-2.3	-1.4	-2.7
Grade 6.2 correlations	.870	.838	.887
Grade 6.2 N's	1695	1684	1684
Grade 6.2 actual means	459.8	482.5	462.8
Grade 6.2 mean of predictions	455.7	481.3	459.7
Grade 6.2 difference (act. - pred.)	4.1	1.2	3.1

TABLE 3. Summary Statistics from the Distribution of Scale Score Differences: Obtained - Predicted.

		CAT-70 Reading Vocabulary	CAT-70 Reading Comprehension	CAT-70 Reading Total
GRADE 1.5 PRI A CAT 1 STANDARD	MEAN	1.13	-.37	.27
	S.D.	21.58	36.63	22.94
	N	431	431	431
GRADE 2.2 PRI A CAT 1 STANDARD	MEAN	-.19	-1.36	-.28
	S.D.	20.34	31.70	20.91
	N	879	879	879
GRADE 2.2 PRI B CAT 1 STANDARD	MEAN	2.14	-6.44	.01
	S.D.	25.79	37.91	28.31
	N	644	644	644
GRADE 2.2 PRI B CAT 1 BLACK	MEAN	-.21	2.51	-1.70
	S.D.	28.25	39.60	30.62
	N	717	717	717
GRADE 3.2 PRI B CAT 1 BLACK	MEAN	2.66	3.39	2.42
	S.D.	26.20	34.49	26.87
	N	843	843	843
GRADE 3.2 PRI B CAT 2 STANDARD	MEAN	-2.90	-2.18	-.49
	S.D.	28.53	37.11	30.14
	N	803	803	803
GRADE 3.2 PRI C CAT 2 STANDARD	MEAN	.19	2.37	.87
	S.D.	30.66	38.25	30.05
	N	532	532	532
GRADE 4.2 PRI C CAT 2 STANDARD	MEAN	-.64	-.11	.48
	S.D.	31.75	37.01	31.71
	N	780	780	780
GRADE 4.2 PRI D CAT 2 STANDARD	MEAN	-2.53	1.66	.53
	S.D.	42.05	44.18	39.09
	N	590	590	590
GRADE 5.2 PRI D CAT 3 STANDARD	MEAN	-2.33	-1.46	-2.71
	S.D.	33.93	40.32	32.48
	N	1390	1390	1390
GRADE 6.2 PRI D CAT 3 STANDARD	MEAN	4.04	1.11	3.17
	S.D.	32.73	39.67	32.21
	N	1697	1697	1697

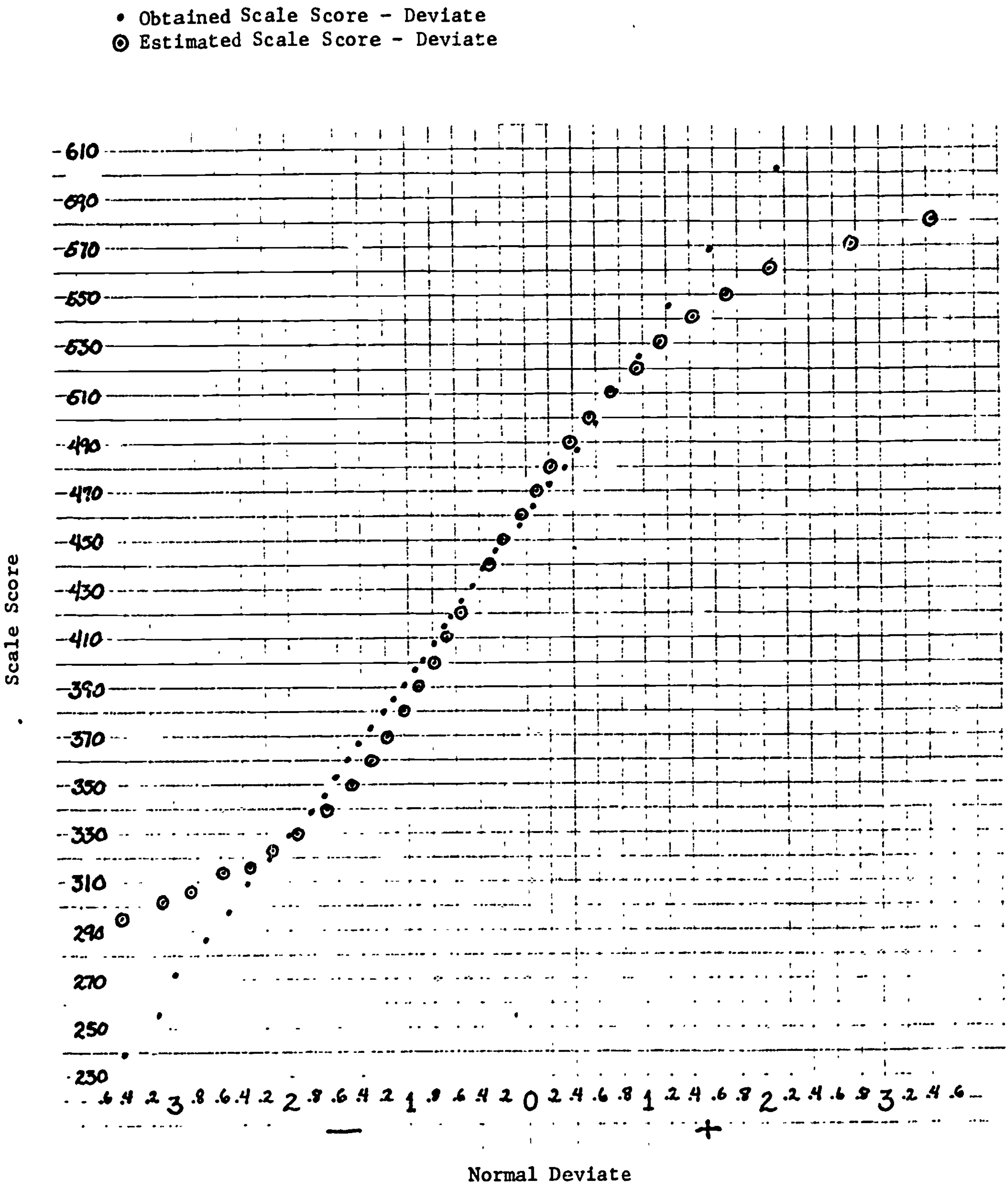

Figure 1. Distribution of Obtained and Predicted Reading Vocabulary Scale Scores for PRI Level D, CAT Level 3, Grade 6.2.