DOCUMENT RESUME

ED 138 613 TM 006 233

AUTHOR Ahn, Unhai R.; Barta, Maryann B.
TITLE An Empirical Comparison of Several Methods for Analyzing and Reporting School Unit Achievement Gains.
PUB DATE [Apr 77]
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Gains; Achievement Tests; Comparative Analysis; Compensatory Education Programs; Elementary Education; Grade Equivalent Scores; *Measurement Techniques; Program Evaluation; Reading Achievement; *Schools; *Scores; *Standardized Tests; Test Interpretation
IDENTIFIERS Normal Curve Equivalent Scores; Percentile Ranks; Residual Scores; Standard Scores

ABSTRACT

School unit achievement gains computed by six different methods are compared. The data include total reading scores from the Metropolitan Achievement Test administered to Title I students in 32 elementary schools. The achievement gains were computed in standard scores, grade equivalents, percentile ranks, residual scores based on individual scores and school means, and normal curve equivalents. The results show that the methods used in analyzing gains on a standardized test affect the outcomes in varying degrees. Correlations ranging from moderate to relatively high indicate that one method can be substituted for another. (Author/RC)

AN EMPIRICAL COMPARISON OF SEVERAL METHODS FOR ANALYZING AND REPORTING SCHOOL UNIT ACHIEVEMENT GAINS

by

Dr. Unhai R. Ahn
and
Maryann B. Barta

Program Evaluation Branch
Cincinnati Public Schools

Presented at annual meeting of the
American Educational Research Association
New York, New York

April, 1977

AN EMPIRICAL COMPARISON OF SEVERAL METHODS FOR ANALYZING AND REPORTING SCHOOL UNIT ACHIEVEMENT GAINS

Dr. Unhai R. Ahn, Maryann B. Barta, Cincinnati Public Schools

Purpose

One of the typical uses of standardized achievement tests has been for measuring present status of an individual or a group of individuals. Recently, however, standardized tests are often being used to measure academic growth. When they are so used, methods for computing gains and ways of aggregating data can lead to different results. Problems encountered in the measurement of academic growth have been discussed by specialists in the area (Harris, 1963). From the practitioner's point of view, it is important to know if the different methods are comparable enough to arrive at the same judgment about students' academic growth and program impact.

The purpose of this paper is to describe several methods of measuring achievement gains, to analyze school unit data using different types of scores, and to compare the different methods for reporting gains. The different types of scores include: standard scores, grade equivalents, percentile ranks, normal curve equivalents (NCE), and residuals.

Background

Different methods for measuring change have been discussed by Lord (1963). In the simplest sense, the observed change (g) is the difference (y-x) between an initial measure (x) and a final measure (y). The positive change in "g" is often called crude gain and negative change is called loss. The estimation of "true gains" was described by Lord

to be an appropriate technique for individuals who are members of a group under consideration. This technique uses multiple regression equations to overcome chance errors of measurement and spurious gains. Dahlke (1971) also pointed out that "true gains" should be considered when measuring reading improvement rather than crude gains.

With current emphasis in education upon accountability and program evaluation, more interest has been shown in measures of achievement gains for a group of individuals or a school unit. Dyer, Linn, and Patton (1969) proposed the use of residuals from multiple regression predictions of school output as indices of school performance. The residuals reported in other studies also showed that they could be used as school performance indices either based on a school unit data or individual scores (Marco, 1974; Gastright, 1975).

In practice, grade equivalent scores are most commonly accepted by teachers and parents in computing performance gains and in performance contracts (Stake, 1971). Grade equivalent scores are somewhat difficult to interpret, especially in the extreme values for each level of a test. For example, if a second grader receives a 4.7 grade equivalent score in science, this score does not mean that the student can comprehend fourth grade science materials. Rather, it means that the student is a high performer in science compared with his/ in the same grade. Realizing these difficulties in interpret ie grade equivalent score, test manufacturers recommend consumers to use percentiles and standard scores instead of grade equivalents for measuring gains (Wrightstone, 1972).

Recently, the U.S. Office of Education (USOE) has developed, in cooperation with the RMC Research Corporation, a new type of standard

score, normal curve equivalent (NCE), to measure gain in ESEA Title I programs across the nation. The normal curve equivalent is a type of standard score which distributes scores equally on the normal curve, with a mean of 50 and a standard deviation of 21 NCE units. Like percentile ranks, NCE's range from 1 to 99 and conversion tables between the two scores are available. (See Appendix A.) The difference between the two, however, is that there are more NCE's at the extremes of the distribution compared to percentiles. Thus, gains in NCE's are larger if the scores change at the very low or very high end of the distribution.

Methods

Data for this study were drawn from 32 elementary schools that participated in the ESEA Title I reading program in Cincinnati during the 1975-76 school year. The students in the sample were the 1,267 third and fourth graders who took a pretest in September, 1975, and a postest in April, 1976. The Metropolitan Achievement Reading Test, Form G, was used. The Primary II level was given to grade three and the Elementary level was given to grade four. Total reading scores were used for both pre and post testings. Standard scores were used for regression analysis and in computing residuals.

Gains in Total Reading were analyzed for each school by six different methods. They are as follows:

1. Mean Standard Score Gain - For each school the mean difference in the standard score was computed. Symbolically, $\bar{G}ss = \bar{Y}ss - \bar{X}ss$, where $\bar{G}ss$ stands for the mean gain in the standard score, $\bar{Y}ss$ is the postest mean in the standard score, and $\bar{X}ss$ is the pretest mean in the standard score.

2. Mean Grade Equivalent Gain - For each school the mean difference in the grade equivalent was calculated. In this case, $\bar{G}ge = \bar{Y}ge - \bar{X}ge$.

3. Percentile Rank Difference Between the Pretest Standing and the Postest Standing - For each School the mean standard score for the pretest as well as the mean standard score for the postest was converted to percentile ranks, and the difference between the two percentile ranks was computed. That is, $\bar{G}p = \bar{Y}p$ (converted from $\bar{Y}ss$) - $\bar{X}p$ (converted from $\bar{X}ss$).

4. Mean Normal Curve Equivalent (NCE) Gain - For each school the mean difference in the NCE was computed. Similarly, $\bar{G}nce = \bar{Y}nce - \bar{X}nce$.

5. Mean School Residual Scores Based on Individual Scores - Individual student postest scores were regressed on individual student pretest scores for about one-fourth of the total students across all schools. A systematic sampling technique was employed to select this group from each grade. With the regression formula known, mean individual residual scores were calculated for each school. That is, Individual Residual = $\frac{1}{N}\sum[Yi - (BXi + \bar{\bar{Y}} - B\bar{\bar{X}})]$, where N is the number of students in the school taking both tests, Yi and Xi are the postest and pretest scores, B is the least squares estimate of the slope for the students across all schools, and $\bar{\bar{Y}}$ and $\bar{\bar{X}}$ are the grand postest and pretest means.

6. School Residual Scores Based on School Means - School postest means were regressed on school pretest means and school residual scores were calculated. This is one of the methods suggested by Dyer (1969) as a measure of school effectiveness. In this case, School Residual = $\bar{Y} - (C\bar{X} + \bar{\bar{Y}}' - C\bar{\bar{X}}')$, where $\bar{Y}$ and $\bar{X}$ are postest and pretest means for the school, C is the least squares estimate of the regression slope of the school's postest means on the school's pretest means, and $\bar{\bar{Y}}'$ and $\bar{\bar{X}}'$ are the unweighted averages of the school postest and pretest means, respectively, across all schools.

Results

The scores computed by the six methods are reported for each school in Tables 1 and 2. Table 1 contains the mean gain and residual scores for grade three in 32 schools; Table 2 contains data for grade four in 29 schools. Three schools, Numbers 3, 15, and 19, did not have ESEA Title I program participants in grade four. Mean gain (+) indicates a positive change on the average from pretest to postest, while mean loss (-) indicates a negative change. Residuals indicate the relative position of the scores; a positive residual indicates above expectation, and a negative residual indicates below expectation (above (+) or below (-) on the prediction line).

The data in these tables show that there is some comparability among the different methods. That is, if one school shows a high gain by one method, it tends to show high gains by the other methods.

Table 1. Gain and Residual Scores in Total Reading, Metropolitan Achievement Test, Grade 3.

School Number	No. of Students	Standard Score	Grade Equiv.	%ile Rank	Normal C.E.	Indiv. Resid.	School Resid.
1	8	9	.8	22	10.0	1.131	1.351
2	18	12	.8	18	17.9	2.218	3.231
3	16	6	.3	5	10.4	-3.173	-2.745
4	17	6	.4	4	9.4	-3.456	-2.769
5	15	10	.7	18	13.1	2.539	.375
6	19	7	.5	12	9.0	-.223	-1.649
7	24	8	.6	10	12.5	-1.377	.231
8	18	11	.7	18	14.9	2.247	1.303
9	23	8	.5	9	10.4	-1.553	-.745
10	20	9	.7	18	11.3	1.350	.351
11	6	8	.6	16	9.7	.450	-.649
12	19	5	.3	4	5.1	-3.573	-3.697
13	11	10	.8	20	14.2	3.076	1.375
14	19	14	1.0	27	20.4	4.335	5.255
15	27	12	.9	24	18.6	3.801	3.303
16	19	8	.6	10	13.2	-1.361	.231
17	29	8	.5	8	12.9	-1.028	-.769
18	20	8	.4	7	11.9	-7.920	-1.745
19	18	10	.6	10	14.5	-2.461	1.183
20	13	8	.4	7	13.5	-2.026	-.793
21	8	11	.6	15	13.3	.647	2.207
22	21	5	.3	3	6.0	-4.654	-3.769
23	24	8	.4	7	11.0	-2.076	-.793
24	23	10	.8	19	13.9	2.392	1.327
25	28	12	.8	21	18.3	3.007	3.255
26	25	10	.6	14	14.1	.687	.279
27	16	6	.4	10	6.9	-1.481	-2.649
28	37	9	.6	14	13.1	.525	.303
29	30	6	.4	9	7.5	-1.987	-2.673
30	13	12	1.0	25	18.2	4.194	3.327
31	30	7	.4	6	11.0	-3.431	-1.769
32	50	7	.5	11	9.1	-.927	-1.673

Table 2. Gain and Residual Scores in Total Reading, Metropolitan Achievement Test, Grade 4.

School Number	No. of Students	Standard Score	Grade Equiv.	%ile Rank	Normal C.E.	Indiv. Resid.	School Resid.
1	9	14	1.5	26	17.1	8.663	8.758
2	19	12	1.0	11	12.2	3.134	3.650
4	7	7	.4	2	5.1	2.947	-1.627
5	18	5	.4	-2	1.2	-.090	-1.580
6	14	11	.8	8	10.9	3.046	1.758
7	24	6	.6	2	1.0	-.654	-.688
8	22	3	.3	0	-1.9	-4.664	-3.519
9	15	6	.5	3	1.7	-2.033	-1.519
10	18	17	1.8	30	19.3	10.160	11.035
11	8	7	.6	4	1.8	-.667	-.519
12	20	5	.4	2	-1.0	-3.136	-2.519
13	11	7	.4	2	5.5	-3.465	-2.181
14	20	10	.8	7	9.2	.912	1.650
16	32	5	.4	1	.1	-2.416	-2.242
17	22	9	.8	8	6.5	.929	1.927
18	35	8	.6	4	3.7	.345	.035
20	26	6	.5	3	1.5	-2.009	-1.519
21	11	8	.7	7	4.6	-.784	1.204
22	32	7	.4	2	2.8	-1.999	-2.073
23	18	4	.4	-2	-1.2	.229	-1.026
24	8	6	.4	0	2.5	-1.334	-1.965
25	19	10	.9	10	9.0	2.956	2.758
26	25	9	.6	4	3.2	-.643	.096
27	30	5	.4	2	.1	-3.298	-2.519
28	41	7	.5	2	1.8	-.984	-1.242
29	27	5	.4	2	1.1	-3.179	-2.519
30	13	6	.6	2	3.5	-.258	-.688
31	23	6	.5	2	1.0	-.752	-.965
32	35	5	.4	0	-.7	-2.390	-1.965

Table 3. Intercorrelations Among the Scores Computed by Different Methods, Grade Three. (N=32)

	Standard Score	Grade Equivalent	%ile Rank	Normal C.E.	Indiv. Resid.	School Resid.
Standard Score	1.00	.91	.85	.92	.83	.97
Grade Equivalent		1.00	.96	.77	.93	.96
Percentile Rank			1.00	.66	.96	.91
NCE				1.00	.68	.85
Individual Residual					1.00	.88
School Residual						1.00

The scores derived from different methods were intercorrelated using Spearman's rank correlation. Tables 3 and 4 contain the intercorrelations among the scores for grades three and four, respectively.

In Table 3 a correlation of ±.349 is significantly different from zero at $\alpha = .05$. It can be noted that all of the correlations are significantly different from zero. Correlations are especially high (.91 or above) for half of the cases. This indicates that the scores are interchangeable for half of the comparisons.

Relatively high correlations are found between grade equivalent and four other types of scores. Relatively low (.77 or below) correlations are found between NCE and percentile rank, individual residual, and grade equivalent. This might suggest that NCE is different than other types of scores and substitutable only for standard score. However, grade equivalent scores seem to empirically represent a very similar rank order of school unit data with other types of scores.

Table 4. Intercorrelations Among the Scores Computed by Different Methods, Grade 4. (N=29)

	Standard Score	Grade Equivalent	%ile Rank	Normal C.E.	Indiv. Resid.	School Resid.
Standard Score	1.00	.87	.90	.94	.74	.84
Grade Equivalent		1.00	.91	.78	.78	.96
Percentile Rank			1.00	.81	.69	.82
NCE				1.00	.72	.76
Individual Residual					1.00	.88
School Residual						1.00

In Table 4 a correlation of $\pm.36$ is significantly different from zero at $\alpha = .05$. All of the correlations on Table 4 indicate a higher value than this.

Relatively high (.91 or above) correlations are found between grade equivalent and school residual, standard score and NCE, and grade equivalent and percentile rank. The correlations between the above three pairs were also high for third grade data. One may say that a similar rank ordering of school unit achievement data can be obtained using either grade equivalent or school residual, standard score or NCE, and grade equivalent or percentile rank.

It may also be noted that NCE's correlate relatively low (.78 or below) with three of the five scores: grade equivalent, individual residual, and school residual. This result was found in the third grade data, too. NCE seems to produce different results than the other scores with the exception of standard score.

Across two samples, school residuals correlate with other types of scores higher than individual residuals do. This is not surprising since school residuals are more directly derived from school means (pre and post mean standard scores) than individual residuals. This may imply that school residuals are better substitutes than individual residuals for mean gain in standard score, grade equivalent, percentile rank, and NCE.

Comparing the third and fourth grade data correlation patterns for both grades are somewhat consistent. However, the correlations for grade three are slightly higher than those for grade four.

Conclusion

This study has not attempted to identify a best method for reporting achievement gains but rather to empirically compare the use of several methods. A case has been made by other authors for the use of residualized gains as a more appropriate indicator of true gain. Grade equivalent gains and mean percentile rank gains have been criticized as inappropriate methods of assessing gain. The former has been criticized because of the curve fitting and equal unit interpolation of grade equivalent months between empirical norm points. The latter has been criticized because of the inappropriate application of individual pupil percentile rank tables to group data. In reality, the percentile distribution of group mean data should be much tighter than that of individual norms, and the use of individual norms on group means should underestimate the actual change in the group. The NCE, although it has the characteristics of a standard score in that it can be averaged, acts in effect as a residual around equipercentile estimation of achievement growth. Unlike the

regression residuals, which are based [illegible] h assumption within some local population, NCE's [illegible] empirically derived non-linear norms found in the percenti[illegible] distributions of the standardized test.

In effect, the use of residualized gains would be the application of local norms, while the use of NCE gains would be the application of an empirical national norm to form residuals. One factor seriously effects the interpretation of NCE gains. "Are the concurrently developed percentile rank norms developed at various norm points unbiased estimates of the longitudinal growth of achievement in the norm population?" This question cannot be answered until large-scale studies of longitudinal achievement growth are conducted on a norm population. The pervasive drop in achievement scores across the country raises a serious question as to the longitudinal validity of these norms.

If the concerns about the validity of the NCE are unfounded, then this method of reporting achievement gains would seem to provide several benefits not found in the other scores. As non-linear residuals from an observed growth curve at each percentile point, they would provide the advantage of representing a nationally valid standard against which the achievement of students at the low and high ends of the distribution could be compared.

The observed differences between the residuals based on school means and those based on the mean residuals of individual data are not surprising, given two factors. First, the sample means are small enough that variability due to error factors alone could effect the individual residual data. Second, the schools are different enough in input achievement levels that the assumption of linear regression could seriously bias the mean of the student residual data.

Two findings are somewhat surprising in this empirical study. Grade equivalent gains and percentile rank gains of mean scores are surprisingly similar to the other measures of gain (NCE excepted). NCE gain is rather surprisingly different from the other measures of gain (standard score gain excepted). The educational and evaluation implications of these similarities and differences need to be studied over a variety of samples and tests so that the evaluative validity of school ranks based on these methods can be reassessed.

APPENDIX A

%ile to Normal Curve Equivalent Conversion Table

%ile	NCE		NCE	%ile	NCE
1	1		42	71	62
2	7		43	72	62
3	10		44	73	63
4	13	39	44	74	64
5	15	40	45	75	64
6	17	41	45	76	65
7	19	42	46	77	66
8	20	43	46	78	66
9	22	44	47	79	67
10	23	45	47	80	68
11	24	46	48	81	68
12	25	47	48	82	69
13	26	48	49	83	70
14	27	49	49	84	71
15	28	50	50	85	72
16	29	51	51	86	73
17	30	52	51	87	74
18	31	53	52	88	75
19	32	54	52	89	76
20	32	55	53	90	77
21	33	56	53	91	78
22	34	57	54	92	80
23	34	58	54	93	81
24	35	59	55	94	83
25	36	60	55	95	85
26	36	61	56	96	87
27	37	62	56	97	90
28	38	63	57	98	93
29	38	64	58	99	99
30	39	65	58		
31	40	66	59		
32	40	67	59		
33	41	68	60		
34	41	69	60		
35	42	70	61		

REFERENCES

Dahlke, A. B. "Predicting True Reading Gains After Remedial Tutoring." In R. Leibert (Ed.), Diagnostic Viewpoints in Reading. Newark, Delaware: International Reading Association, 1971.

Dyer, H. S. "T[illegible]riteria of Professiona[illegible]ccountab[illegible]lity in the Schools o[illegible] Phi Delta Kappan, 5[illegible] (Dec., 1970) 206-211.

Dyer, H. S., Linn, R. L., Patton, M. J. "A Comparison of Four Methods of Obtaining Discrepancy Measures Based on Observed and Predicted School System Means on Achievement Tests." American Educational Research Journal, 1969; 6, 591-605.

"Title I Evaluation Models Set By USOE." Education U.S.A. July 19, 1976, Vol. 18, No. 47, p. 269.

Gastright, J. F. "The Effects of Adding Student Mobility and Demographic Variables on the Stability and Comparability of Multiple Regression Indices of School Performance: After Dyer's Models." Doctor[illegible]issertation, University of Cincinnati, 1975.

Harris, C. W. (Ed[illegible] Problems i[illegible]ring Change. Madi[illegible]University of Wisconsin Pres[illegible]1963.

Lord, F. M. "El[illegible]tary Models for [illegible]asuring Change." i[illegible]. Harris (Ed.), Problems i[illegible]M[illegible]ring Change. Madison: University of [illegible]sin Press, 1963.

Marco, G. L. "A Comparison of Selected School Effectiveness Measures Based on Longitudinal Data." Journal of Educational Measurement, 1974, 11, 225-34.

O'Connor, E. "Extending Classical Test Theory to the Measurement of Change." Review of Educational Research, 42 (Winter, 1972), 73-97.

Stake, R. E. "Testing Hazards in Performance Contracting." Phi Delta Kappan, 1971, 52, [illegible]-89.

Winer, B. [illegible]ical Principles [illegible]Experimental Design. New York: McGraw-Hill, [illegible].

Wrightstone, J. [illegible]gan, T. P., Abbott, M. M. "Accountability in Education and Associate[illegible]easurement Problems." Test Service Notebook 33. New York: Harcourt Brace Jovanovich, 1972.