DOCUMENT RESUME

ED 222 567 TM 820 725

AUTHOR Powers, Stephen; And Others
TITLE A Test of the Equipercentile Hypothesis of the TIERS Norm-Referenced Model.
INSTITUTION Tucson Unified School District, Ariz.
PUB DATE Sep 82
NOTE 25p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Gains; Compensatory Education; *Hypothesis Testing; Models; National Norms; *Norm Referenced Tests; Pretests Posttests; Program Effectiveness; Reading Achievement; Secondary Education; *Validity
IDENTIFIERS *Equipercentile Assumption; RMC Models; *Title I Evaluation and Reporting System

ABSTRACT

The validity of the equipercentile hypothesis of the Title I Evaluation and Reporting System (TIERS) norm-referenced evaluation model was examined. The California Achievement Test, Reading, was administered as a pretest and posttest to 3,224 seventh and ninth grade students. The equipercentile hypothesis predicts that the posttest percentile status would be the same as the pretest percentile status for students not receiving special education programs. Students' gains at 10 different achievement levels were evaluated employing the norm-referenced model. The findings contradicted the equipercentile hypothesis. There was a clear pattern of large gains for students not receiving any special educational instruction. (Author/CM)

A TEST OF THE EQUIPERCENTILE HYPOTHESIS OF THE TIERS NORM-REFERENCED MODEL[1]

Stephen Powers

Tucson Unified School District

Helen Slaughter

Tucson Unified School District

Cheryl Helmick

Tucson Unified School District

Legal and Research Services

Tucson Unified School District

September 1982

Running head: Equipercentile

Abstract

The validity of the equipercentile hypothesis of the TIERS norm referenced evaluation model was examined using 3,224 seventh and ninth grade students. The California Achievement Test, Reading, was administered as a pretest and a posttest. The equipercentile hypothesis predicts that the posttest percentile status would be the same as the pretest percentile status for students not receiving special educational programs. Students' gains at ten different achievement levels were evaluated employing the norm referenced model. Confidence interval procedures were used. The findings contradicted the equipercentile hypothesis. There was a clear pattern of large gains for students not receiving any special educational instruction.

A Test of the Equipercentile Hypothesis of the TIERS Norm-Referenced Model[1]

Estimating the achievement gains of students between pre- and post-tests for the purpose of evaluating the effectiveness of educational programs is perhaps one of the most widely used evaluation models in American education. Called the norm-referenced model or Model A in the federally-mandated Title I Evaluation and Reporting System (TIERS), this model is used to evaluate the progress of approximately 99 percent of students participating in Title I--the largest federally-funded program for educationally disadvantaged students (Linn, Dunbar, Harnisch, & Hastings, 1982).

The norm-referenced model is based on a strong assumption--the equi-percentile assumption--which specifies that without special supplementary programs such as those funded through Title I, students' posttest percentile status would remain the same as their pretest percentile status. The equipercentile assumption was defined by Tallmadge and Wood (1976, p. 4) as follows:

> When tests with national norms are used, the no-treatment expectation is found by determining the percentile status of the treatment group at pretest time. It is assumed that, without the Title I treatment, the status of the group at posttest time would be the same as it was at pretest time.

Therefore, within the purview of the norm-referenced model, increases in percentile rank reflect gains due to programmatic effect. Perhaps because

the equipercentile assumption is so intuitively appealing, there has been only limited research testing the validity of this key assumption of the TIERS norm-referenced model.

It has been noted that the equipercentile assumption has minimal empirical support (Horst, Tallmadge, & Wood, 1975) and theoretical support (Echternacht, 1978). Kaskowitz and Norwood (1977) found a tendency for the equipercentile curve to underestimate expected posttest scores for extremely low pretest scores and to overestimate posttest scores for extremely high pretest scores. Van Hove, Coleman and Karweit (1970) using cross-sectional data reported considerable changes in percentile ranks across time. Echternacht (1978), using Monte Carlo techniques to simulate test and learning behavior, tentatively concluded that Model A overestimated the treatment effect.

Tallmadge (1982) examined the norm-referenced model employing data files from the Sustaining Effects Study (SES) and the national norming of the California Achievement Tests (CAT). A major focus of his study was on the norm-referenced gain estimates of low achieving students in Grades 2, 4, and 6 from fall to spring. Although gain estimates varied from -.34 NCE to 2.62 NCE for different size Local Education Agencies and from -2.21 (city) to 8.33 (large city), Tallmadge reported that overall there was a positive bias of about 1 NCE for Title I groups.

While Tallmadge's study (1982) is enlightening, there were some limitations to the inferences that could be drawn about norm-referenced gains because (1) the SES analysis employed an on-level selection test and posttest and a below-level pretest, (2) in the CAT analyses three to four

combinations of forms and levels of the CAT were used for the pretest, (3) in the CAT analyses, norm-referenced gains were calculated for groups which formed a substantial portion of the norms they were compared to, and (4) the correlations between the selection test and pretest and posttest were not calculated.

The following are the rules for implementation of the norm-referenced model (Model A1) as specified in Tallmadge and Wood (1976, pp. 40-41): (1) a nationally normed achievement test should be administered as a pretest and posttest, (2) whenever possible, the same level and form of the test should be administered as a pretest and posttest, (3) participants must not be chosen on the basis of their pretest scores, (4) participants should be tested on a level of the test appropriate to their functional level, and (5) all testing should be accomplished within two weeks of the empirical norming dates. However, Tallmadge and Wood (1976) added that interpolated norms could be used: "By interpolating between the surrounding data points, testing times can be extended from September 8 to October 22 and March 26 to May 7." (p. 41)

The purpose of the present study was to test the equipercentile hypothesis using a sample of students from schools which did not participate in special supplementary educational programs. Some of the research hypotheses which will be considered in this study are: Will the equipercentile hypothesis hold at ten different levels of achievement? If the equipercentile hypothesis does not hold, will larger biases occur with the more extreme groups? Will biases occur when a selection test is administered two years before the pretest? Essentially, the present study is a

test of the following null hypothesis: if the equipercentile hypothesis is valid and the requirements of the norm-referenced model are adhered to, students not receiving special supplementary educational programs will not be expected to show gains in achievement over time relative to national norms.

Method

Sample

The sample consisted of 3,224 seventh and ninth grade students attending nine junior high schools and seven high schools in a metropolitan school district in the Southwest with an enrollment of approximately 51,000 students. All students with complete data sets (selection test, pretest, and posttest) were included in the sample. None of these schools participated in projects funded through Title I of the Elementary and Secondary Education Act (ESEA) or the Emergency School Aid Act (ESAA). The sample included 48% males and 52% females. The ethnic composition of the sample was 1% American Indian, 4% Black, 2% Asian, 17% Hispanic, and 75% Anglo (non-Hispanic Caucasians). The ethnic composition of the national norm group consisted of 15% Blacks, 10% Hispanics and 75% Others.

Instrumentation

The selection tests which were administered two years before the pretests were the following: (1) seventh grade students were tested during the week of October 5, 1978 with the Comprehensive Tests of Basic Skills (CTBS), 1975 Edition, Level 2, Form S, Total Reading Test, (2) ninth grade students were tested the week of September 25, 1978 with the California Achievement Test (CAT), 1977 Edition, Level 17, Form C, Total Reading Test.

Seventh and ninth grade students were pre- and posttested during the 1980-81 school year with the same form and level of the CAT, 1977 Edition, Form C, Total Reading Test. Seventh grade students were administered Level 17 and ninth grade students, Level 18 of the CAT. Both groups were pretested during the first three weeks of September 1980 and posttested during the week of April 20, 1981. Since the pretest was administered during the first three weeks of September and not within two weeks of the norming dates, appropriate CAT interpolated norms were used (CTB/McGraw-Hill, 1979). Use of interpolated norms was the only instance where the present study varied from the requirements of the norm-referenced model.

Research Design

The confidence interval model was selected for this study rather than the hypothesis testing model which has often been criticized by statisticians (Kish, 1959; Savage, 1957; Tukey, 1954; Yates, 1951). Statistical estimation appeared to be more appropriate than tests of significance which would allow only the rejection of the null hypothesis. Furthermore, confidence interval procedures tell the researcher "how much faith he can place in his estimates and they indicate how much the N needs to be increased to raise the precision of estimates by particular amounts" (Nunnally, 1960, p. 647). In summary, the confidence interval approach appeared to be more informative than the hypothesis testing model (Linn, Note 1).

Students were grouped into ten 10-percent intervals according to percentiles of the selection test. These ten 10-percent intervals ranged from the 1-10 percentile interval to the 91-99 percentile interval. The smallest group consisted of 48 students within the 1-10 percentile interval of the seventh grade and the largest group was 335 in the 91-99 percentile interval of the ninth grade. It was expected that selection with a test other than the pretest would reduce the regression effect operating on the pre- and posttest scores.

Percentile scores of the pretest and posttest were converted to Normal Curve Equivalent (NCE) units. The NCE scale is a normalized standard score scale ranging from 1 to 99 with a mean of 50 and a standard deviation of 21.06. Norm-referenced gain estimates were calculated by subtracting the group's fall pretest NCE mean from the spring posttest NCE mean. For each of the ten groups in the seventh and ninth grade, these gain estimates were calculated with accompanying 95% confidence intervals. One can utilize a confidence interval as a significance test since establishing a confidence interval implies a test of significance (Edwards, 1954). For example, if the hypothesized population value falls outside the 95% confidence interval, then a test of significance with alpha at .05 would result in the rejection of the null hypothesis.

According to the equipercentile hypothesis the parameter of interest is zero since it is hypothesized that there will be no gain for students who are not receiving special educational programs. The 95% confidence interval is constructed so that there is 95% probability of including the value of the parameter between its limits.

The most serious treatment to a pre- posttest research design when interest is focused on low or high achieving students is the regression effect, the so-called "ubiquitous phenomenon" (Campbell & Stanley, 1963, p. 11). Linn (1981, p. 94) succinctly explained the regression effect:

> When students are selected according to their standing on some indicator of achievement . . . the group will regress toward the mean on any correlated measure of achievement obtained at a later point in time. The lower the correlation between the measure used for selecting participants and the subsequent measure, the greater the regression toward the mean.

Linn (1981) also noted that the pretest and the posttest scores will regress toward the population mean even though a separate selection measure is used. The magnitude of the regression effect would depend on the correlation between the selection measure and subsequent measures. Glass (as cited in Linn, 1981, p. 94) noted that the regression effect for the pretest will not equal the regression effect for the posttest. It could be expected that the posttest would regress more toward the mean than the pretest because the selection test would correlate less with the posttest than it does with the pretest (Linn, 1981).

Results

The equipercentile hypothesis that the status of a "no-treatment" group at posttest time would be the same as it was at pretest time was not supported by the findings of this study. Contrary to the expectations of the equipercentile hypothesis, posttest NCE means were consistently higher than pretest NCE means. The differences between pre- and posttest NCE

means were large in many cases (for example 8.26 and 7.23 NCEs) and fifteen of the twenty confidence intervals failed to include the expected parameter of zero. One may conclude that the percentile status at the posttest time was higher than the percentile status at pretest in most of the cases.

In each ten percentile interval of the selection test, seventh grade subgroups exhibited NCE mean gains from pretest to posttest. The mean NCE gain for all seventh grade students was 3.50. Mean gains of the subgroups ranged from .06 (1-10 percentile interval) to 8.29 (21-30 percentile interval). Seventh grade low achieving students tended to show greater gains than higher achieving students with mean gains of the subgroups generally declining linearly from the 11-20 percentile interval to the 91-99 percentile interval. Eight of the ten subgroups gains were statistically significant beyond the .001 level (Table 1). A visual presentation of data showing mean gains with 95% confidence intervals plotted as a function of the 10-percent intervals of the selection test is found in Figure 1.

Insert Table 1 about here

Insert Figure 1 about here

Ninth grade students in each ten percent subgroup exhibited mean NCE gains ranging from 1.55 to 2.70. Neither higher nor lower achieving

students showed greater gains. Seven of the mean gains were significant at the .05 level (Table 2). The ninth grade data are presented visually in Figure 2 with mean gains and 95% confidence intervals plotted as a function of the selection test 10-percent intervals. Overall the mean NCE gain was 2.14 for grade 9 students.

Insert Table 2 about here

Insert Figure 2 about here

Overall mean NCEs indicate the seventh and ninth grade achievement was above the national norms. The mean seventh grade NCE for the selection test was 59.24 (SD = 18.59), for the pretest was 58.61 (SD = 19.20), and for the posttest was 62.11 (SD = 18.11). The correlation between the selection test and the pretest was .86 and between the selection test and the posttest was .85.

Ninth grade results were similar to the seventh grade results. The mean ninth grade NCE for the selection test was 59.45 (SD = 19.14), for the pretest was 58.89 (SD = 18.33) and for the posttest was 61.03 (SD = 18.86). The correlation between the selection test and the pretest was .86, and between the selection test and the posttest was .84.

The correlations between the selection test and the pre- and posttests appeared high considering there was a two-year period between the selection test and the pretest. The distribution of scores of both the seventh and ninth grade students was somewhat skewed, indicating a large proportion of

high achieving students. For example, of the seventh grade students, 10% scored in stanines 1-3 and 37% scored in stanines 7-9. This was not unexpected as low economic level--low achieving schools were not included in the analysis.

Often students are selected for Title I because they scored in stanines 1-3 on some selection test. In an additional analysis three groups were formed based on stanines 1-3, 4-6, and 7-9 of the selection test to increase the generalizability of results to Title I programs. Furthermore, in the previous analyses, the subgroups of 10-percent intervals had widely varying standard deviations, lower selection, pretest/posttest correlations, and lower reliabilities. By selecting students from a larger interval, it was hoped to approximate more the distribution of scores in Title I evaluations.

Students in each of the three subgroups of the seventh and ninth grades demonstrated mean gains. Of special relevance to Title I evaluation, a mean gain of 4.52 was exhibited by seventh grade students in stanines 1-3 and a mean gain of 1.86 for ninth grade students in stanines 1-3 (Table 3).

Insert Table 3 about here

In summary, the equipercentile hypothesis did not appear to hold across ten different ability levels, no clear pattern of greater biases occurred with extreme groups, and large biases occurred in spite of the fact that the selection test was administered two years before student selection.

Discussion

The findings of this study contradict the no-treatment expectations of the equipercentile hypothesis. Furthermore, these results are especially convincing because they show a clear pattern for students' gains to be overestimated. These findings are consistent with the regression hypothesis that selection of students on a test other than the pretest will not completely eliminate the regression effects.

These findings are consistent with those of Echternacht (1978) who found that Model A will overestimate gains. Kaskowitz and Norwood's (1977) findings are not completely consistent with these findings although they did find a tendency to overestimate gains for high pretest scoring students. Because Kaskowitz and Norwood used cross-sectional norms, their findings could be due to differences in the different norming samples. The present finding of a consistent overestimation at each ability level is especially convincing because the same students, tested on the same form and level of the CAT, were compared with the longitudinal norms of the CAT. Moreover, since students were not selected on the pretest, the overestimation of gains is in agreement with the regression hypothesis that the posttest will regress more than the pretest. Tallmadge (1982) found a positive bias of about 1 NCE for low achieving students in the elementary school grades. The present study found an even greater bias in the norm-referenced model than did Tallmadge.

Generalization of these findings to Title I students' gains is not without some limitations. The present study included only seventh and

ninth grade students. Students were selected into achievement groups on the basis of a test administered two years before the pretest. Finally, the present study employed interpolated norms to adjust for the pretesting before the time of empirical norms.

The equipercentile assumption is the key assumption of the norm-referenced model. Researchers have found a tendency for a positive bias in this assumption. The present study is a straightforward test of the equipercentile hypothesis in which a pattern of overestimation of gains has been found. These gains have been very large indeed, providing empirical evidence seriously questioning the validity of the equipercentile assumption. These findings also strongly suggest that research employing the norm-referenced model will find gains where none exist.

Reference Note

1. Linn, R. L. Personal Communication, 1981.

Footnote

[1]The authors would like to express their appreciation to Robert L. Linn who made helpful suggestions during the initial phases of this study, to Darrell L. Sabers for his technical advice and constructive comments during the preparation of this paper, and to Gary Estes for critiquing an earlier version of this paper. However, the opinions and conclusions expressed herein are those of the authors.

Authors

Stephen Powers. Address: Legal and Research Services Department, Tucson Unified School District, 1010 E. Tenth St., Tucson, Az. 85719. Title: Research Specialist. Degrees: B.S. Northern Arizona University, M.A., M.Ed., Ph.D. University of Arizona. Specialization: Program Evaluation.

Helen Slaughter. Address: Legal and Research Services Department, Tucson Unified School District, 1010 E. Tenth St., Tucson, Az. 85719. Title: Research Evaluator. Degrees: B.S. Northern Arizona University, M.Ed., Ed.D. University of Arizona. Specialization: Program Evaluation and Ethnographic Research.

Cheryl Helmick. Address: Legal and Research Services Department, Tucson Unified School District, 1010 E. Tenth St., Tucson, Az. 85719. Title: Research Specialist. Degree: B.A. University of Texas at Austin. Specialization: Program Evaluation.

Table 1

Mean Gains and 95 Percent Confidence Intervals for Ten 10-Percent Intervals of Seventh Grade Students

Interval	N	Mean Gain	Standard Deviation	95 Percent Confidence Interval
1-10	48	.60	12.25	-2.95, 4.15
11-20	75	7.23	10.98	4.71, 9.76
21-30	55	8.29	13.29	4.71, 11.87
31-40	85	5.03	10.27	2.81, 7.25
41-50	119	4.67	8.74	3.08, 6.26
51-60	148	4.10	7.88	2.82, 5.38
61-70	199	3.51	6.94	2.54, 4.48
71-80	205	2.37	6.70	1.45, 3.29
81-90	192	3.23	8.26	2.06, 4.04
91-99	201	1.11	8.71	- .10, 2.32
TOTAL	1327	3.50	8.86	

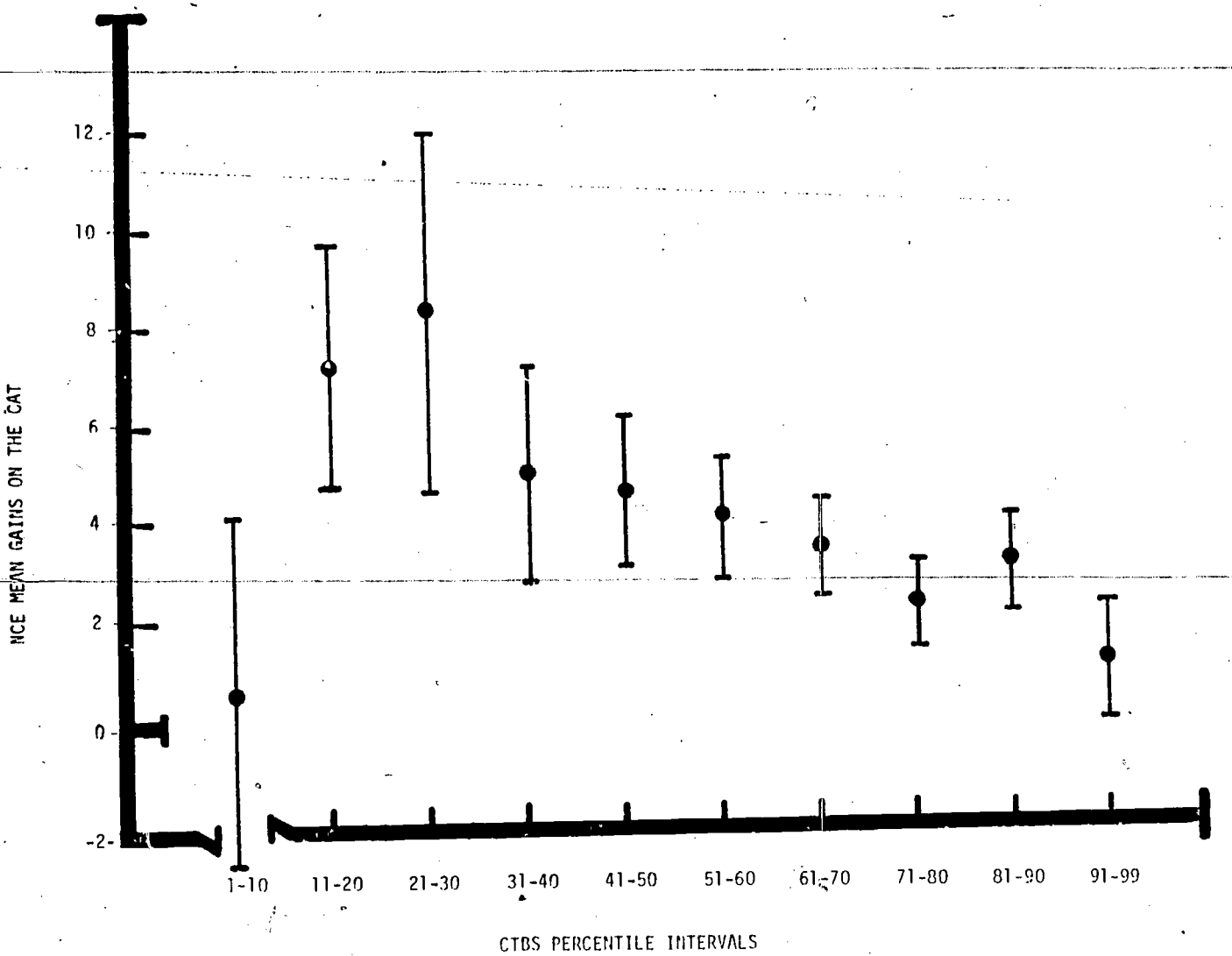

Figure 1. Seventh grade students NCE mean gains on the California Achievement Test with 95 percent Confidence Intervals vs. Comprehensive Tests of Basic Skills

Table 2

Mean Gains and 95 Percent Confidence Intervals for Ten 10-Percent Intervals of Ninth Grade Students

Interval	N	Mean Gain	Standard Deviation	95 Percent Confidence Internal
1-10	59	2.26	10.58	- .49, 5.01
11-20	85	2.42	12.31	- .24, 5.06
21-30	112	1.55	9.29	- .19, 3.29
31-40	184	1.97	8.25	.77, 3.17
41-50	177	1.63	7.90	.45, 2.81
51-60	178	1.95	9.10	.60, 3.30
61-70	225	2.42	7.64	1.42, 3.42
71-80	249	1.89	8.38	.89, 2.89
81-90	293	2.70	8.59	1.71, 3.69
91-99	335	2.19	9.68	1.15, 3.23
TOTAL	1897	2.14	8.91	

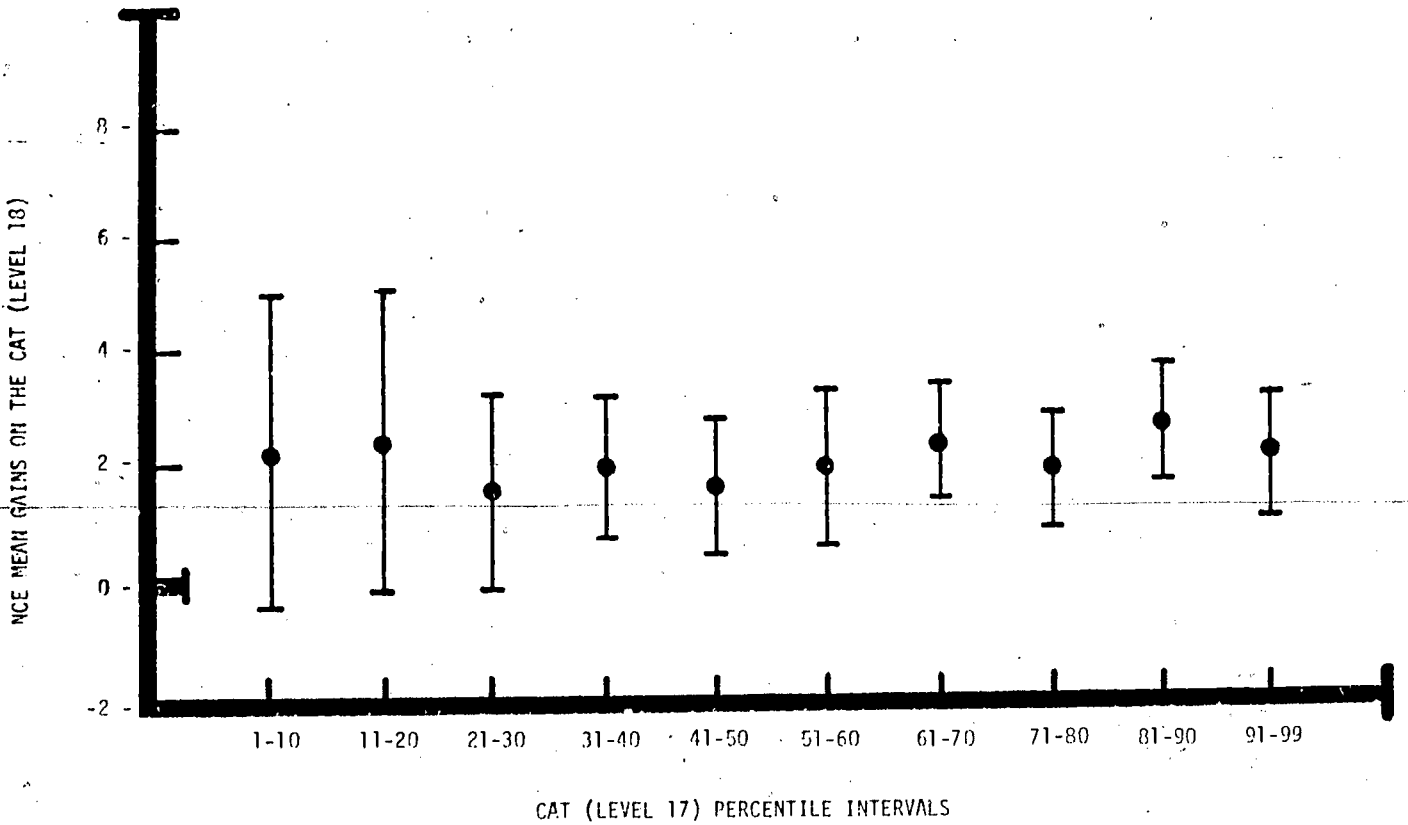

Figure 2. Ninth grade student NCE mean gains on the California Achievement Test (Level 18) with 95 percent Confidence Intervals vs. California Achievement Test (Level 17)

Table 3

Mean Gains for Seventh and Ninth Grade Students

Grade	Stanine Interval	N	Mean Gain	Standard Deviation
Seventh	1-3	131	4.52	11.83
	4-6	710	4.18	8.49
	7-9	486	2.24	8.30
Total	1-9	1327	3.50	8.86
Ninth	1-3	171	1.86	11.07
	4-6	978	1.99	8.36
	7-9	748	2.39	9.06
Total	1-9	1897	2.14	8.91

REFERENCES

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

CTB/McGraw-Hill California Achievement Tests: Interpolated Norms. Monterey, Calif.: Author, 1979.

Echternacht, G. The use of different models in the ESEA Title I evaluation system. Paper presented at the annual meetings of the American Educational Research Association, Toronto, April 1978.

Edwards, A. L. Statistical methods for the behavioral sciences. New York: Rinehart, 1954.

Horst, D. P. Tallmadge, G. K., & Wood, C. T. A Practical Guide to Measuring Project Impact on Student Achievement (No. 1 in a series of monographs on evaluation in education). Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1975.

Kaskowitz, D. H. & Norwood, C. R. A study of the norm-referenced procedure for evaluating project effectiveness as applied to the evaluation of project information packages. Research Memorandum. Menlo Park, CA; Stanford Research Institute, January 1977.

Kish, L. Some statistical problems in research design. American Sociological Review, 1959, 24, 328-338.

Linn, R. L. Measuring pretest-postest preformance changes. In R. A. Berk (Ed.), Educational Evaluation Methodology: The State of the Art. Baltimore: The Johns Hopkins University Press, 1981.

Linn, R. L., Dunbar, S. B., Harnisch, D. L., & Hastings, C. N. The validity of the Title I evaluation and reporting system. Paper presented at the annual meeting of the American Educational Research Association, New York City, 1982.

Nunnally, J. The place of statistics in psychology. Educational and Psychological Measurement, 1960, 20, 641-650.

Savage, R. J. Nonparametric statistics. Journal of the American Statistical Association, 1957, 52, 332-333.

Tallmadge, G. K. An empirical assessment of norm-referenced evaluation methodology. Journal of Educational Measurement, 1982, 19, 97-112.

Tallmadge, G. K. & Wood, C. T. User's Guide: ESEA Title I Evaluation and Reporting System (Revised). Mountain View, CA: RMC Research Corporation, December 1976.

Tukey, J. W. Unsolved problems of experimental statistics. Journal of the American Statistical Association, 1954, 49, 710.

Van Hove, E., Coleman, J. S., Rabben, K., & Karweit, N. School performance: New York, Los Angeles, Chicago, Philadelphia, Detroit, Baltimore. Unpublished manuscript, Baltimore, October 1970.

Yates, F. The influence of Statistical Methods for Research Workers on the development of the science of statistics. Journal of the American Statistical Association, 1951, 46, 32-33.