DOCUMENT RESUME

ED 350 313 TM 019 008

AUTHOR	Lang, Michael H.; And Others
TITLE	Does the Mean Score Mask Poor Delivery of Educational Services in School Effectiveness Ratings?
PUB DATE	Apr 92
NOTE	30p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; Classification; Comparative Analysis; Delivery Systems; Educational Quality; Effective Schools Research; *Elementary Schools; Grade 3; *Low Achievement; Primary Education; Public Schools; Regression (Statistics); *School Effectiveness; *Scores; State Surveys
IDENTIFIERS	*Educational Indicators; Kappa Coefficient; Louisiana; *Mean (Statistics); Residuals (Statistics); Service Delivery Assessment

ABSTRACT

This study investigated whether mean scores in school effectiveness ratings were masking poor delivery of educational services to low achievers in a sample of 242 Louisiana public elementary schools accounting for over 18,000 third graders tested in 1989. Ten separate multiple regression models, each producing studentized residuals used as school effectiveness indicators (SEIs), were used. The SEIs classified schools as effective, average, or ineffective. Appropriate cross-classification results were analyzed, and each comparison was tested with the kappa z-test. The kappa z-tests were significant beyond the 0.05 level, and magnitude measures were generally high to moderately consistent for mean quartile comparisons. The most inconsistent classifications were between effective and average ratings. None of the SEI sets demonstrated a significant relationship with the independent variables in the regression models. Findings consequently indicate that few schools were classified as average on mean-based SEIs that were rated as ineffective on lower quartile based SEIs. Little mean-masking of lower quartile achievement is present. Six tables present study data, and 20 references are included. (SLD)

DOES THE MEAN SCORE MASK POOR DELIVERY OF EDUCATIONAL SERVICES IN SCHOOL EFFECTIVENESS RATINGS

by Michael H. Lang
Charles Teddlie
Jeffery Oescher

Preseented at the April 1992 NCME Annual Meeting in San Francisco

DOES THE MEAN SCORE MASK POOR DELIVERY OF EDUCATIONAL SERVICES IN SCHOOL EFFECTIVENESS RATINGS

by Michael H. Lang
Charles Teddlie
Jeffery Oescher

Since 1969, researchers have been employing regression procedures to evaluate schools on effectiveness by attempting to control for the influence of hard-to-change variables on standardized achievement test scores (Lang, 1991; Mandeville & Anderson, 1987). Originally those hard to change variables were measures of previous learning (Dyer, Linn, & Patton, 1969); later many researchers became interested in controlling for variables of home environment (Wimpelberg, Teddlie, & Stringfield, 1991). Research has vacillated between the two independent variables (IVs) and has sometimes employed measures of both in the regression model (Mandeville & Anderson, 1987).

Regardless of the IV employed, many researchers have criticized the use of mean-scores as the bases for creating dependent variables (DVs). More specifically, that criticism was that the use of mean scores on achievement tests as school-level DVs a masked the inadequate delivery of educational services to those who were in need (Purkey & Smith, 1983; Good & Weinstein, 1986).

The major interest with regards to this study was ,hether setting the level of aggregation to that of the school in regression analysis had classified some schools as effective when major portions of their population were being ineffectively served. That is, was the employment of the school mean scores on achievement tests as DVs in regression analysis masking poor delivery of educational services to lower achievers? This question addresses the issue of equity verses efficiency dilemma in delivery of educational services discussed by Wimpelberg, Teddlie, & Springfield (1989) . In particular, the issue regards the efficiency of the overall mean score verses the equity of subgroup analysis where alternate levels of aggregate performance were considered (e.g., lower-quartile scores).

Equity versus efficiency is an important theoretical issue in school effectiveness research. Mean masking was contextually attached to the equity/efficiency issue in the evolution of effective school research as traced by Wimpelberg et al. (1989). In their study, the researchers noted that the early stages of the effective schools movement were characterized by efforts toward proving that the lower socioeconomic strata could be educated. They termed those stages as the "equity phase" of school effectiveness research. Following that phase, was what they named the "efficiency phase" in which research was broadened to study other groups served by education.

Similarly, mean scores, which appear to have dominated test score data for regression analyses, were related to the efficiency phase in that mean scores were the most efficient representations of school performances. Other aggregate scores only represented points along the range of scores and did not demonstrate control for background data as well as mean scores did (Abalos, Jolly, & Johnson, 1985; Marco, 1974; and O'Connor, 1972).

On the other hand, results from high-risk subgroups (e.g., the lower 50% in achievement) can provide data on how educational services are being delivered to those children with whom the early effective school researchers, such as Edmonds and Frederiksen (1979), were concerned. For this study, the issue regarded whether schools measured in an efficient manner (i.e., mean-based SEIs) were considered as effective when they were measured with concerns to equity (e.g., lower-quartile-based SEIs).

Equity was made an issue in school effectiveness research with Edmonds (1979). In addressing the educational needs of the urban poor, Edmonds (1979) explained his stand on equity:

> "By equity I mean a simple sense of fairness in the distribution of the primary goods and services that characterize our social order. At issue is the efficacy of a minimum level of goods and services to which we are all entitled. Some of us, rightly, have more goods and services than others, and my sense of equity is not disturbed by that fact. Others of us have almost no goods and access to only the most wretched services, and that deeply offends my simple sense of fairness and violates the standards of equity by which I judge our social order." (p.15)

LITERATURE REVIEW

Research on techniques of isolating the effectiveness of individual schools has evolved over a two-decade period since the Dyer et al. (1969) study had attempted to control for student background variables with the regression model. Within that time frame, researchers conducted numerous studies on effective schools, employing various techniques of which the regression model was most frequently used (Lang, 1991).

Even though the regression model has been the one of preference over the past two decades, there have been numerous criticisms regarding the use of school averages as either DVs or as IVs. Those criticisms generally focused on the concern that mean scores have been masking ineffective delivery of educational services to low income and/or low achieving students (Geske & Teddlie, 1990; Good & Brophy, 1986; Purkey & Smith, 1983; Rowan et al., 1983). Furthermore, the effective school concept is essentially multi-level--student, class, and school (Sirotnik & Burstein, 1985). The residual model considers only the uppermost level, the school.

Good and Weinstein (1986) questioned the use of mean school level data in the effective school research which preceded their report, noting,

> "Student averages can be misleading. Although the literature focuses on average difference between schools in attainment, there is ample evidence that a good deal of variation occurs within schools.... Thus we need to move from average effects to effects in individual classrooms and for different kinds of children." (p.1093).

School averages may have been masking within-school variation in previous school evaluation projects.

Concerned with the use of average data from which to determine school effectiveness, Rowan, Bossert, and Dwyer (1983) indicated that employing aggregated data ignored important variations within schools. They noted, "Even within curriculum areas and at a single grade level, schools may not be uniformly effective for all types of students." (p. 27).

Such was the concern of Edmonds (1979) when he set the requirement for effectiveness to be that a school provide low-income children the same minimum level of basic skills mastery as that provided middle-income children. In taking a stand for equity in education, he rated the nation's schools which taught low-income children as "dismal failures" (page 15).

In consideration of the Edmonds equity issue, Geske and Teddlie (1990) suggested conducting a separate analysis for the students scoring in the lowest quartile in addition to the mean-based regression used in effective school analyses. The authors suggested that separate lower-quartile analysis "enables researchers to study school effectiveness simultaneously from the equity and efficiency perspectives." (p. 212).

In a review of effective schools research, only two projects attempting to disaggregate data using regression analysis were found (Marco, 1974, and Dyer et al., 1969). Marco compared five regression variations including two methods of disaggregating data to compute residuals. Dyer et al. analyzed within-school regression slopes to determine the relative effectiveness of a given school for its high and low achievers.

Concerned with the effect of a single indicator of school effectiveness, Marco (1974) employed three indicators, each representing a different within school performance level. He explained the design for disaggregating the data in his study in the following manner:

> "Since a single school effectiveness index may be misleading, three indices were calculated for each school using the within-school regression models. Reference points were selected to represent low-, middle-, and high-scoring students. These points were the mean individual pretest score across all schools and points one standard deviation above and below the mean.... The school effectiveness indices were the regression estimates of the mean posttest scores at these three reference points." (p. 228).

The IVs and DVs in the Marco study were obtained from the fall and spring administrations of the Primary II Metropolitan Reading Achievement Test, forms F and G. Marco correlated the various indices with 30 different variables; however, the relationships of primary interest here were the SEI correlations between all combinations of methods and the SEI-IV correlations for each method.

In terms of their relationships to the IV, the low-scoring based residuals and the middle-scoring based residuals both moderately correlated to the pretest scores ($\underline{r}$ = .41), indicating a substantial amount of IV influence remaining in the residuals. However, the high-scoring based residuals demonstrated a similar relationship to the IV as did the student-based residuals ($\underline{r}$ = .26 and .28 respectively). On the other hand, the correlation of school-based residuals to the IV was zero, indicating a total absence of IV influence.

Marco suggested several explanations for the non-zero correlations of residuals with the IV, including a lack of controls and a limited sample size (70 schools and 3769 students). Though he included controls for prior learning in his study, Marco did not attempt to control for SES variables.

Perhaps, the results of disaggregation on achievement test scores were confounded by also using achievement test scores as the IV in the study. That is, employing subgroups which were segregated for aggregation purposes on a similar basis as they were to be controlled may have neutralized the controls.

Marco concluded that the five methods employed in his study varied enough in results that they should not be used singularly or interchangeably. "The school effectiveness indices for the initially low- and high-scoring students appear to give unique information and raised doubts about using a single index to measure the effectiveness of a school for a given group of students." (Marco, 1974, p. 233).

In the Dyer et al. (1969) study, slope analysis was a secondary part of the overall analyses. The purpose of the slope analysis was to determine if differential delivery of educational services was present within schools.

Their study found that the test of homogeneity was statistically significant at or below the .001 level for all measures; that is, the subgroup slopes were significantly different in every case. Such results suggested that effectiveness was different for the two groups across all measures.

Though disaggregated data appeared to have demonstrated less student background control than aggregated data, disaggregated data in both the Marco (1974) and Dyer et al. (1969) studies provided a more indepth view of how effectively the schools in their research were delivering educational services. In search for a tighter fitting model with better controls for student background data, researchers may have been straying from the original issue--the education of all children, including lower achievers.

Sirotnik and Burstein (1985) discussed this issue in their expository article on multi-level educational research, stating, "To be sure, using statistics (other than the mean) can present some rather sticky analytical issues. Nevertheless, if averages do not fit the constructs being measured, then there is no point pretending that they do." (p. 177). What they were saying is that if a given statistic did not serve the purposes of its project, then that statistic should have not been employed, regardless of how well it performed.

RESEARCH QUESTION

This study researched the effect that varying the DVs had on the consistency of residual-based school effectiveness ratings. In particular, the major interest of this study was whether mean-score aggregations of achievement test results produced the same school effectiveness ratings as did lower-quartile representations of school performance on the same test results.

> The question that study raised was whether mean-score based DVs had produced the same results as lower-quartile based DVs in classifying effective schools.

For the research question, the initial hypothesis was that some degree of relationship existed between the methods of evaluating school effectiveness. That is, the relationship between two sets of school classifications was significantly more than what would have been expected by chance alone. If the hypothesis was accepted, then the two school-rating sets were considered somewhat consistent with each other.

More importantly, the second hypothesis was that the magnitude of the relationship between the two sets of school effectiveness classifications was sufficient enough that schools can be expected to be correctly classified regardless of which DV was selected. That is, does sufficient consistency exist between the two classification sets to warrant the continued use of mean scores in computing regression based School Effectiveness Indices (SEIs). The alternate hypothesis was that the mean-score was masking unequal delivery of educational services to the lower quartile subgroup.

RESEARCH DESIGN

This study analyzed consistency in which two sets of school effectiveness rating models classified schools along various achievement tests. The study employed multiple regression techniques to obtain the base-line data from which to classify schools along effectiveness. The design of each rating model differed by varying the point of regression (i.e., the DV) from the school's mean score to its lower-quartile score in each model for each achievement test.

The regression models were used to compute school-level residuals while controlling for identified student background variables (IVs). Those residuals were used as school effectiveness indices (SEIs) to classify schools along three levels of effectiveness (Mandeville & Anderson, 1987).

All SEIs were computed by standardizing the school residuals in each regression model. The method of standardizing residuals in this study was studentizing the residuals along a t-score distribution. From these studentized residuals, each school was classified as either effective, average, or ineffective. These classifications were used as the basis for the study's subsequent consistency analyses.

The school-level SEIs were computed twice for each test -- once using the mean score as the point of regression, next using the lower quartile as the point of regression. To obtain SEIs for each school, the regression point for every school on each achievement test was predicted from the linear relationship of school-aggregated student background variables and school-aggregated test scores across the data set. The predicted regression point was then subtracted from the actual regression point to produce a raw school residual for each test; the raw residual was then studentized, producing the SEI from which a given school was classified.

The SEI represented whether that school had performed higher or lower than expected. If its performance was substantially higher than expected (i.e., a high positive SEI), the school was classified as effective. If its performance was substantially lower than expected (i.e., a low negative SEI), the school was classified as ineffective. If its performance reached neither extreme, then the school was considered average.

Three SES variables were employed as IVs in the regression models. Those SES variables included teacher-reported data on level of parent-education (percent of mothers who were college graduates) and parent-employment (percent of fathers who were white collar workers), and student-reported data on school lunch status (percent of students on paid-lunch status). The SES data had originally been collected in categorical format on the student level during the spring test administration. It was aggregated to the school level for this study.

Regression procedures required separate procedures for each DV. In conducting separate procedures, the IVs were held constant across all models in order to determine the effect on consistency while the DVs were manipulated. There were ten different DVs used in this study.

The DVs in the study were the school mean scores and lower-quartile scores on criterion-referenced tests (CRTs) for language arts and mathematics, and on norm-referenced tests (NRTs) for reading, language, and mathematics. The NRTs and the CRTs chosen were those grade appropriate tests which were administered to public schools throughout the state of Louisiana in the spring of 1989.

The measurement instruments used to compute the DVs were the Level 13, Form E, California Achievement Tests (CAT-13) and the Grade 3 Louisiana Educational Assessment Program tests (LEAP-3). The CAT-13 is an NRT instrument; the LEAP-3 is a CRT instrument.

The CAT-13 had been normed for use with third grade students; the LEAP-3 had been designed to measure third-grade language and mathematics skills as stipulated in the Louisiana curriculum guides for those subjects. The LEAP-3 is a grade-level test, not a minimum skills test (Louisiana Department of Education, 1989).

The LEAP-3 is administered annually to all Louisiana public school students in the third grade as a measure of how well individual students, schools, districts, and the state are addressing the grade-level curricula in language arts and mathematics. The CAT-13 is administered annually in many public school districts in Louisiana as a measure of how well third-grader performances relate to a nationally designed norm. Some school districts restrict the testing of the CAT-13 to partial populations, apparently as an aid in placement into remedial and special education classes, though most districts employing the CAT-13 measure their total population.

Nearly 250 Louisiana elementary schools whose third grade populations were tested with both the NRT and the CRT in 1989 formed the study sample. That sample was taken from a larger sample used in a recent study (Oescher et al., 1989) compiled from scores for third grade students in the state's public schools who had taken both NRT and CRT tests. The unit of analysis Oescher et al. study was the student; the unit of analysis for this study was the school.

The final sample was a reduced one reflecting the removal of inappropriate data for school-level analyses. Such data included the following cases: (1) districts which had not attempted to test their total populations with the NRT, (2) schools whose demographic data were in question, (3) schools which had been poorly matched on CRT and NRT scores, and (4) students who had been absent for the administration of the CRTs and had been assigned a zero score in that data set by default.

The number of schools represented in the final data set were 242, accounting for more than 18,000 students. The percent black was 52.9%, the percent white was 44.4%, and the percent of other ethnicity was 2.7%. The proportions of the final sample in terms of gender were 50.5% male and 49.5% female. With regard to ethnicity, the final sample did not reflect the state's population. The black population was oversampled; the white population was undersampled.

To determine the level of school effectiveness, a classification criteria was established for the studentized residuals: +/-0.674 standard error units (se). Those schools with SEIs beyond than +0.674 se for any DV were classified as "effective" for that DV; those schools with SEIs beyond -0.674 se were classified as "ineffective" for that DV. Those schools with SEIs from +0.674 se to -0.674 se for any DV were classified as "average" for that DV.

The reasoning behind the choice of those points were (1) that the outlier status of beyond +/-0.674 se should have been moderate enough as to have minimized the regression effect on subsequent studies of the same schools, (2) that half of the schools were expected to be classified as average, assuming the SEIs to be normally distributed (Glass & Hopkins, 1984), and (3) that the categorical distributions were similar in size (25%-50%-25%) so as to minimize the influence of chance agreement (Reynolds, 1970).

The design for the consistency analyses of the study's comparisons crossed the results of the mean-based SEIs with that of the quartile-based SEIs for the same assessment instrument in five separate contingency tables:

(1) classifications derived from mean scores crossed with those derived from lower-quartile scores on CRT language arts test;

(2) classifications derived from mean scores crossed with those derived from lower-quartile scores on CRT mathematics test;

(3) classifications derived from mean scores crossed with those derived from lower-quartile scores on NRT reading test;

(4) classifications derived from mean scores crossed with those derived from lower-quartile scores on NRT language test;

(5) classifications derived from mean scores crossed with those derived from lower-quartile scores on NRT mathematics test;

All five contingency tables were 3-by-3 in design for each level of school effectiveness: effective, average, and ineffective. The purpose of the contingency tables was to compare the results of the two classification models.

The comparisons were tested to determine if significant consistency existed. The consistency of the school effectiveness classifications were measured for each issue using the kappa _z_-test of agreement to determine if varying the DV significantly affected classification decisions. Additionally, magnitude measures of agreement were computed for each comparison to determine the degree of consistency.

The most straight-forward measure of agreement is the unweighted agreement ratio. The unweighted agreement ratio served in this study as a measure of absolute agreement. It is the percent of classifications with which two models concur; it is the sum of the diagonal cells divided by the total units in the analysis. With a possible range from zero to one, the ratio gauges the numerical proportion of identical classifications to the total classifications.

This statistic was employed as the measure of absolute agreement. With regards to this type of magnitude measure, all agreements were absolute, there were no partial agreements; hence, all disagreements were also absolute.

The weighted agreement ratio is a variation in which the elements in off-diagonal cells are weighted inversely as to their degree of disagreement. Regarding a three-level contingency table, neither agreement or disagreement is absolute with the weighted agreement ratio. Perfect agreement cells were weighted with a 1.0, and the perfect disagreement cells were weighted with a 0; the other cells which represent partial disagreement (or agreement) were weighted with 0.5.

A third variation of percent agreement is the kappa coefficient. That statistic controls for chance agreement expected from the distribution of the data. It employs the table's row and column totals (marginals) in determining expected agreement. This study employed a weighted kappa coefficient which was an extension of the weighed agreement ratio. The general range of kappa is +1.0 for perfect agreement to 0 where the agreement ratio equals expected chance agreement. Kappa values are negative where the agreement ratio is less than what is expected by chance (Reynolds, 1977).

For significance testing, the kappa _z_-statistic was chosen as the measure of consistency because it was not as sensitive to the sample size as measures of association and because it controlled for chance consistency. A significant _z_-test means that the two classification distributions demonstrate some agreement; an insignificant test means that the two distributions are independent of one another--there is no significant agreement beyond what would be expected by chance. The _z_-statistic is computed by dividing the kappa coefficient by its standard deviation (Reynolds, 1977).

Finally, the SEIs from each regression model were correlated with the IVs to determine if the regression procedure adequately controlled for the influence of the IVs in the analyses as was the concern in the Marco (1974) study. Those correlations were t-tested to determine if they significantly deviated from zero. Zero correlations were indicative that the SEIs were independent of the IVs.

FINDINGS

The comparisons of mean-based and quartile-based school effectiveness classifications demonstrated significant agreement for all five pair-wise results. In addition, the degrees of magnitude as measured by the kappa coefficient and the weighted agreement ratio were substantial (See Tables 1, 2, 3, 4, & 5).

Table 1
Contingency Table Comparison of School Classifications by CRT Language Arts Mean & Lower Quartile SEIs

Quartile Results:	Ineffective		Average		Effective		Row Total	
	n	%	n	%	n	%	n	%
Mean-Based Results:								
Ineffective	52	(21.5)	11	(4.6)	0	(0.0)	63	(26.0)
Average	6	(2.5)	104	(43.0)	12	(5.0)	122	(50.4)
Effective	0	(0.0)	11	(4.6)	46	(19.0)	57	(23.6)
Column Total	58	(24.0)	126	(52.1)	58	(24.0)	242	(100.0)

Statistical Results

Comparison Levels:	Effective Average Ineffective	Average Ineffective	Average Effective
Statistics:			
Kappa Coefficient	.776	.794	.700
Kappa Z-Statistic	2.99^{**}	2.25^{***}	2.11^{***}
Unwghted Agreement Ratio	.835	----	----
Weighted Agreement Ratio	.917	.905	.867

Note: Decision points between average and the other categories are +/-.674 se on the studentized residual distribution.

* probability < .001
** probability < .01
*** probability < .05

According to the kappa coefficients, the general finding was that one out of every four schools had discrepant classifications, that is with chance agreement controlled. The weighted agreement ratio indicated that approximately only one out of eleven schools had discrepant classifications. However, further analyses revealed that most of the discrepancies were between average and effective classifications, meaning that the means were primarily masking ineffective delivery of services to upper-level subgroups in some schools.

Regarding absolute agreement, the general finding was that approximately one of every six schools were inconsistently classified (See Tables 1-5). Again, most of the discrepancies were found between the average and effective classifications.

Table 2
Contingency Table Comparison of School Classifications by CRT Mathematics Mean & Lower Quartile SEIs

Quartile Results:	Ineffective		Average		Effective		Row Total	
	n	%	n	%	n	%	n	%
Mean-Based Results:								
Ineffective	51	(21.1)	8	(3.3)	0	(0.0)	59	(24.4)
Average	8	(3.3)	100	(41.3)	16	(6.6)	124	(51.2)
Effective	0	(0.0)	14	(5.8)	45	(18.6)	59	(24.4)
Column Total	59	(24.4)	122	(50.4)	61	(25.2)	242	(100.0)

Statistical Results

Comparison Levels:	Effective Average Ineffective	Average Ineffective	Average Effective
Statistics:			
Kappa Coefficient	.744	.790	.620
Kappa Z-Statistic	2.76^{**}	2.35^{***}	1.84^{***}
Unwghted Agreement Ratio	.835	----	----
Weighted Agreement Ratio	.917	.904	.829

Note: Decision points between average and the other categories are +/-.674 se on the studentized residual distribution.

* probability < .001
** probability < .01
*** probability < .05

With regard to data distributions, the NRT scores were generally negatively skewed within schools, and the CRT scores were both negatively and positively skewed within schools. Where tests were skewed, the within-school indices generally fell between +/- 1.0 for the CRTs and between 0 and -1.0 for the NRTs. Both ranges indicated that the degree of skew was slight.

Though the cross school skewness indices provided clues to the nature of the within school distributions, it is the within school skewness that may influence the classifications of individual schools. Such within school distributions have apparently been influencing the consistency indices for NRT comparisons and may have been influencing some indices for CRT comparisons.

Table 3
Contingency Table Comparison of School Classifications by NRT Language Mean & Lower Quartile SEIs

Quartile Results:	Ineffective		Average		Effective		Row Total	
	n	%	n	%	n	%	n	%
Mean-Based Results:								
Ineffective	50	(20.7)	11	(4.5)	0	(0.0)	61	(25.2)
Average	8	(3.3)	102	(42.1)	11	(4.5)	121	(50.0)
Effective	0	(0.0)	11	(4.5)	49	(20.2)	60	(24.8)
Column Total	58	(24.0)	124	(51.2)	60	(24.8)	242	(100.0)

Statistical Results

Comparison Levels:	Effective Average Ineffective	Average Ineffective	Average Effective
Statistics:			
Kappa Coefficient	.772	.755	.719
Kappa Z-Statistic	2.95^{**}	2.30^{***}	2.15^{***}
Unwghted Agreement Ratio	.831	----	----
Weighted Agreement Ratio	.915	.889	.872

Note: Decision points between average and the other categories are +/-.674 se on the studentized residual distribution.

* probability < .001
** probability < .01
*** probability < .05

Table 4
Contingency Table Comparison of School Classifications
by NRT Reading Mean & Lower Quartile SEIs

Quartile Results:	Ineffective		Average		Effective		Row Total	
	n	%	n	%	n	%	n	%
Mean-Based Results:								
Ineffective	48	(19.8)	12	(5.0)	0	(0.0)	60	(24.8)
Average	8	(3.3)	109	(45.0)	8	(3.3)	125	(51.7)
Effective	0	(0.0)	14	(5.8)	43	(17.8)	57	(23.5)
Column Total	56	(23.1)	135	(55.8)	51	(21.1)	242	(100.0)

Statistical Results

Comparison Levels:	Effective Average Ineffective	Average Ineffective	Average Effective
Statistics:			
Kappa Coefficient	.756	.744	.705
Kappa Z-Statistic	3.09*	2.32***	2.27***
Unwghted Agreement Ratio	.826	----	----
Weighted Agreement Ratio	.913	.887	.874

Note: Decision points between average and the other categories are +/-.674 se on the studentized residual distribution.

* probability < .001
** probability < .01
*** probability < .05

To the degree which negative skewness was present for a given test, the mean and lower quartile scores converged toward on another. A school with a negative skewed score distribution would have a mean score (1) lower than its central grouping of scores, (2) lower than its median, and (3) closer to its lower quartile scores (Glass & Hopkins, 1984). With a symmetric distribution, a given school would have a mean score in line with its median and central grouping.

Hence, a negative skew would be more apt to result in similar classifications along mean and lower quartile regression models because of the relative proximity of the two reference points (mean and lower quartile). Where substantial negative skewness is present, one can expect little mean masking of lower quartile achievement regardless of differential delivery of educational services.

Table 5
Contingency Table Comparison of School Classifications by NRT Mathematics Mean & Lower Quartile SEIs

Quartile Results:	Ineffective		Average		Effective		Row Total	
	n	%	n	%	n	%	n	%
Mean-Based Results:								
Ineffective	49	(20.3)	11	(4.5)	0	(0.0)	60	(24.8)
Average	8	(3.3)	107	(44.2)	11	(4.5)	126	(52.1)
Effective	0	(0.0)	13	(5.4)	43	(17.8)	56	(23.1)
Column Total	57	(23.6)	131	(54.1)	54	(22.3)	242	(100.0)

Statistical Results

Comparison Levels:	Effective Average Ineffective	Average Ineffective	Average Effective
Statistics:			
Kappa Coefficient	.753	.756	.681
Kappa Z-Statistic	2.97^{**}	2.33^{**}	2.12^{**}
Unwghted Agreement Ratio	.822	----	----
Weighted Agreement Ratio	.911	.891	.862

Note: Decision points between average and the other categories are +/-.674 se on the studentized residual distribution.

* probability < .001
** probability < .01
*** probability < .05

Regarding a related issue, a question exists as to whether alternate sources of SEIs, such as the lower-quartile scores, would also demonstrate zero or near zero correlation with IVs (Sirotnik & Burstein, 1985; Marco, 1974). This study has employed SES variables as IVs; the Marco (1974) study employed previous test scores as IVs.

The findings of this study were that the IVs did not significantly correlate with the SEIs (α=.05). That means the regression models adequately controlled the influence of IVs on the SEIs which were employed to classify schools (See Table 6).

Table 6

Correlation Matrix of Independent Variables with School Effectiveness Indices

Independent Variables: Fathers	% of Stdnts	% of Mothers	% of
Collar	Paid Lunch	College Grad	White
Sch. Effect. Indices:			
CRT Math. Mean	+.0006	+.0016	+.0008
CRT Lang. Mean	+.0004	+.0018	+.0009
NRT Math. Mean	+.0004	+.0011	+.0005
NRT Lang. Mean	-.0005	+.0007	-.0006
NRT Read. Mean	-.0002	+.0013	-.0000
CRT Math. Quartile	+.0004	+.0015	+.0005
CRT Lang. Quartile	+.0009	+.0021	+.0014
NRT Math. Quartile	+.0004	+.0008	+.0004
NRT Lang. Quartile	-.0004	+.0012	-.0002
NRT Read. Quartile	+.0003	+.0016	+.0005

Note: None of the correlations were significantly different from zero for p<.05. Those findings indicated that SES was controlled in each regression model.

CONCLUSIONS

The question that this study raised was whether the mean was masking poor delivery of educational services to lower achievers. The study compared school effectiveness classifications based on two points of regression: the mean and the lower quartile. Given the current concern with the state of education in this nation, such a question focuses attention towards that subgroup which the equity advocates have historically considered under-educated.

The crossing of the classification results from mean and lower quartile-based regression models demonstrated significant agreement along effective school classifications for each test considered. The degree of magnitude found in each comparison was moderate.

However, mean masking was primarily found in the average rating of effective delivery of educational services to the lower quartile population. Very little mean masking was found in the average rating of ineffective delivery of such services to that group.

Subsequent two-level analyses revealed that most of the consistency in each three-level comparison was in the average-ineffective decisions. Hence, there was less evidence of mean masking of ineffective delivery of educational services to the lower quartile than there was of effective delivery of services to that same group.

The degree of negative skewness within school scores appeared not to be substantial. Therefore, the influence of the skewed distribution on consistency appears to be subtle. That conclusion is supported by reviewing the pattern of agreement ratios and kappa coefficients in Tables 1-5. One would expect that the consistency measures for the NRTs to be higher than the CRTs because of the greater degree of negative skew. However, no such pattern was demonstrated. Perhaps for this study, the degree of equivalence that educational services were delivered to the lower quartile students had more of an influence on consistency than did the distributions of scores.

Ineffective delivery of education of the lower quartile group, the crux of this study, has been the major concern of those educators and researchers who pioneered the equity phase of the school effectiveness movement (Edmonds, 1979; Edmonds & Frederiksen, 1979; Wimpelberg et al., 1989). Additionally, mean masking of ineffective education for that group has been a major focus of criticism during the efficiency phase of the movement (Good & Weinstein, 1986; Rowan et al., 1983; and Purkey & Smith, 1983). Though the concern in literature was substantial, the evidence from this study indicated there was little problem with mean masking to the lower quartile group.

The following conclusions have been drawn regarding this study: (1) that employing SES variables as IVs are not be as problematic as employing previous test scores where the lower quartile is employed as the point of regression; (2) that the inconsistencies found in this study between mean and quartile models, though limited, are substantial enough to warrant separate analyses for each situation; and (3) that employing the lower quartile in the regression model will hold schools accountable for those individuals from that part of the school population targeted for need by Edmonds.

IMPLICATIONS

The mean masking concern of earlier researchers should not be taken lightly. It may be beneficial for school evaluation programs not to employ mean scores by default. Instead, such programs can conduct their own quartile analysis to determine if substantial mean masking is present in their situation. Local consistency studies are generally feasible wherever school mean scores are being employed. That is, whenever a mean score can be computed from raw score data, a lower-quartile score can also be computed.

In an absolute sense, mean-masking of ineffective delivery of educational services to even a small number of schools may be intolerable to decision makers. In those cases, DVs derived from lower quartile scores could be employed.

Future effective school studies can explore modeling possibilities with disaggregated data which can provide a broader-based view of school effectiveness than does the mean-score. In the case of quartile analysis, the median score represents the middle score for the total school uninfluenced by outliers; the upper quartile score represents the median for the upper half of scores; the lower quartile represents the median for the lower half.

REFERENCES

Abalos, J., Jolly, S.J., & Johnson R. (1985). Statistical methods for selecting merit schools. Paper presented at the annual meeting of the American Educational Research Association, Chicago. ED261097.

Dyer, H.S., Linn, R.L., Patton, M.J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 6(4), 591-605.

Edmonds, R.R. (1979). Effective schools for the urban poor. Educational Leadership, 37, 15-24.

Edmonds, R.R., Frederiksen, J.R. (1979). Search for effective schools: The identification and analysis of city schools that are instructionally effective for poor children. ED 170396.

Geske, T, & Teddlie, C. (1990). Organizational productivity of schools. In P. Reyes (ED) Teachers and their workplace: Commitment, performance and productivity. Newbury: Sage Publications.

Glass, G.V., & Hopkins, K.D. (1984). Statistical Methods in Education and Psychology. Englewood Cliffs, NJ: Prentice-Hall.

Good, T.L., & Brophy, J.E. (1986). School effects. Third Handbook of Research on Teaching, Ed. M. Whitrock, New York: Macmillan.

Good, T.L., & Weinstein, R.S. (1986). Schools make a difference. American Psychologist, 41(10), 1090-1097.

Lang, M.H. (1991). Effective School Status: A Methodological Study of Classification Consistency. Dissertation, Louisiana State University, Baton Rouge.

Levine, D.U., & Lezotte, L.W. (1990). Unusually Effective Schools. Austin: The National Center for Effective Schools Research & Development.

Louisiana Department of Education (1989). Louisiana Educational Assessment Program Annual Program Report, 1988-89 School Year. Baton Rouge.

Mandeville, G.K., & Anderson, L.W. (1987). The stability of school effectiveness indices across grade levels and subject areas. Journal of Educational Measurement, 24(3), 203-216.

Marco, G.L. (1974). A comparison of selected school effectiveness measures based on longitudinal data. Journal of Educational Measurement, 11(4), 225-233.

O'Connor, E.F. Jr. (1972). Extending classical test theory to the measurement of change. Review of Educational Research, 42(1), 73-97.

Oescher, J., Paradise, L.V., & Kirby, P.C. (1989). Comparison study of Grade 3 norm-referenced and criterion-referenced test results. Report prepared for the Louisiana Department of Education, Baton Rouge.

Purkey, S.C., & Smith, M.S. (1983). Effective schools: A review. The Elementary School Journal, 83(4), 427-452.

Reynolds, H.T. (1977). The Analysis of Cross Classifications. New York: Free Press.

Rowan, B., Bossert, S.T., & Dwyer, D.C. (1983). Research on effective schools: a cautionary note. Educational Researcher, 12 (4), 24-31.

Sirotnik, K.A., & Burstein, L. (1985). Measurement and statistical issues in multilevel research on schooling. Educational Administration Quarterly, 21(3), 169-185.

Wimpelberg, R.K., Teddlie, C., & Stringfield, S. (1989). Sensitivity to context: The past and future of effective schools research. Educational Administration Quarterly, 25(1), 82-107.

DOES THE MEAN SCORE MASK POOR DELIVERY OF EDUCATIONAL SERVICES IN SCHOOL EFFECTIVENESS RATINGS

by Michael H. Lang
Charles Teddlie
Jeffery Oescher

Presented at the April 1992 NCME Annual Meeting in San Francisco.

ABSTRACT

This study investigated whether mean scores were masking poor delivery of educational services to low achievers in such evaluations.

The sample included 242 Louisiana public elementary schools (18,000 third graders tested in 1989). The study employed ten separate multiple regression models, each producing studentized residuals used as school effectiveness indicators (SEIs). The independent variables for all models were student's free lunch status, mother's educational level, and father's employment level. The dependent variables were school mean and lower quartile scores for CRT language arts and mathematics tests, and NRT reading, language, and mathematics tests.

The study used SEIs to classify schools as effective, average, or ineffective. It classified each school according to ten different models using +/-.674 *se* as the post hoc criteria.

The study separately analyzed appropriate cross classification results: (1) CRT language arts mean & lower quartile, (2) CRT mathematics mean & lower quartile, (3) NRT language mean & lower quartile, (4) NRT reading mean & lower quartile, (5) NRT mathematics mean & lower quartile.

The study tested each comparison with the kappa *z*-test; it measured agreement with the weighted kappa coefficient (chance-controlled agreement), the weighted agreement ratio (adjusted agreement), and the unweighted agreement ratio (absolute agreement).

The study found the kappa-*z* tests significant beyond the .05 level. It found that magnitude measures were generally high-moderately consistent for mean-quartile comparisons. It found that most inconsistent classifications were between effective and average ratings. It also found that all SEI sets demonstrated no significant relationship with the independent variables in the regression models.

Hence, the findings indicate that there were few schools classified as average on mean based SEIs that were classified as ineffective on lower-quartile based SEIs. The study concludes that findings indicate that little mean-masking of lower quartile achievement is present.