Pergamon

Studies in Educational Evaluation, Vol. 20, pp. 113–145, 1994

0191–491X/94 $24.00

0191-491X(94)E0008-4

EFFECTIVENESS INDICES: A "VALUE ADDED" APPROACH TO MEASURING SCHOOL EFFECT

William J. Webster, Robert L. Mendro and Ted O. Almaguer

Dallas Independent School District, Dallas, TX, U.S.A.

One of the most important aspects of administering a large urban school district is the identification of schools that are unusually effective with the students they serve. Once effective schools are identified, detailed studies can be launched to determine the concomitants of their effectiveness and those concomitants can be replicated in similar environments across the district. Researchers investigating effective schools have consistently identified five to seven factors that are correlated with improved school achievement (Good & Brophy, 1986; Purkey & Smith, 1983). These factors include a sense of mission (Brookover & Lezotte, 1979; Ohio Department of Education, 1981); strong building leadership (Edmonds, 1982; Shoemaker & Fraser, 1981); high expectations for students and staff (Clark, Lotto & McCarthy, 1980: Eubanks and Levine, 1983); frequent monitoring of student progress (Edmonds, 1982); a positive, orderly learning climate (Edmonds, 1982; Shoemaker & Fraser, 1981); sufficient opportunity for learning (Brookover & Lezotte, 1979; Edmonds, 1982; MacKenzie, 1983); and parent/community involvement (Ohio Department of Education, 1981; Stedman, 1985).

While the contribution of the effective schools movement has been substantive and such research will obviously be continued, there are a number of related questions, concerns, and criticisms emerging in educational discourse. Specifically, concerns posed by a number of investigators include that the development of new techniques for evaluating school effectiveness has not kept pace with the increased and continuing interest (Felter, 1989; Webster & Olson, 1988); that most studies that have been done have focused upon narrow educational outcomes, typically norm-referenced achievement tests (Rowan, Bossert & Dwyer, 1983; Stedman, 1987, 1988); that the research has been primarily limited to elementary schools in urban systems with large populations of disadvantaged youth (Clark, Lotto & McCarthy, 1980; Farrar, Neufeld & Miles, 1983; Firestone & Herriott, 1982; Rowan, Bossert & Dwyer, 1983); and that most of the research attempting to associate school effects with student learning is correlational meaning that causation has not generally been established (D'Amico, 1982; Neufeld, Farrar & Miles, 1983; Purkey & Smith, 1983). If research related to effective schools is to be advanced, new techniques for identifying and evaluating effective schools need to be developed (Saka, 1989). Inherent in this task are two complex issues: (1) a better

application include its relative simplicity of application and interpretation, its robustness, and the fact that general methods of structuring complex regression equations to include combinations of categorical and continuous variables and their interactions are quite straightforward (Aiken & West, 1991; Cohen, 1968; Cohen & Cohen, 1975; Darlington, 1990).

Another method of incorporating a large number of input, process, and outcome variables into an equation is canonical correlation (Van de Geer, 1971). The canonical correlation model calls for the use of canonical correlation to establish a linear combination of dependent and independent variables from which average deviation levels can be established for each school. In essence, the average deviation from the first canonical variate or first several variates for the dependent variables would be determined for each school. Schools would then be ranked on this deviation to produce the school effectiveness indices. In essence, for multilevel and student-level variables, each level of each variable would be standardized and the average value of the standardized levels computed. Average standardized scores for each level would be combined to produce a total standard score for the variable. Then, all dependent and independent variables would be used to establish the first canonical correlation for the data set. The values of the first canonical variate for the linear combination of dependent variables would be determined for each school. Deviations from the district dependent variable portion of the first canonical variate would then be computed for each school. The deviations from the canonical variate would be used to rank schools for the school effectiveness indices. As in the other models, the deviations would be standardized.

A third appropriate statistical method would be time-series analysis (Fuller, 1976; Nelson, 1973). Time-series analysis calls for the use of equations to establish predicted scores for each dependent variable that has sufficient historical levels of data available. The average deviation from the predicted levels of the dependent variables would be determined for each school and schools would be ranked on this deviation. For multilevel and student-level variables, each level of each variable would be standardized and the average value of the standardized levels computed. Average standardized scores for each level would be combined to produce a total standard score for the variable. Then, a predicted value for each dependent variable would be computed. Deviations from the predicted values would then be computed for each school. The deviations would be averaged and the average deviation used to rank schools for the school effectiveness indices. Once again, the deviations would be standardized.

Finally, hierarchical linear modeling was considered. Hierarchical linear modeling is an application of multiple regression analysis that provides different equations at different levels of observation. Education data are often hierarchical. Students are grouped into classes that are grouped into schools that are grouped into districts, etc. Hierarchical linear modeling takes this hierarchical structure into account and thus makes it possible to incorporate variables from all levels. Thus equations might be developed at the class, school, and district levels (Bryk & Raudenbush, 1992).

In theory building, one must be concerned with the basic assumptions inherent in traditional linear model analysis, those being linearity (although regression equations can be adjusted to reflect non-linear relationships), normality, homoscedasticity, and independence. In applications where the entire population is involved and one is using multiple regression analysis as a descriptive technique, the requirements of independence of individual observations as well as of normality have no practical impact. In fact,

individual observations are being used to determine school effect so that non-independence of student observations within classroom and school is what is being sought. Linearity and homoscedasticity are still important statistical issues and must be dealt with.

The Accountability Task Force, given these various options, chose multiple regression analysis. Canonical correlation was not chosen because the Task Force wanted to be able to differentially weight variables by perceived level of importance. Time series analysis was not the technique of choice because it required at least three years of longitudinal data and, in an urban school district, population mobility is such that these requirements would have had major impact on the degrees of freedom of the equations. Hierarchical linear modeling was not chosen because of degrees of freedom problems in the within school equations and the fact that the authors were not concerned with explaining school effects, merely with validly identifying effective schools. For this reason, path analysis, another explanatory technique, was not considered.

This paper discusses several distinct multiple regression models and the issues and problems related to their application to the practical problem of identifying effective schools. The results obtained from the application of each of these models are discussed. The first model was implemented in 1984, the second in 1992.

Method

Effective Schools Methodology - 1984

When faced with the challenge of identifying effective schools in 1984, it was decided to identify effective schools in terms of student achievement in the basic skill areas of reading, mathematics, and language usage (Webster & Olson, 1988). While many other desirable goals and outcomes of public education could have been easily enumerated, at that time heightened public awareness was focused on standardized test scores. A school was not considered successful unless its student test scores improved.

School Effects. In these early studies a school's effectiveness was associated with exceptional student achievement, defined as measured test performance above or below that which would be expected if a school did no more or no less than simply maintain students' previous rates of achievement growth. In general it is reasonable to expect students to continue to achieve at a given rate. When a school's population of students departed markedly from its own pre-established trend or from the more general trend of similarly achieving students throughout the district, this departure was attributed to a school effect. The problem of measuring a school's effect, then, becomes one of establishing the school's students' rates of achievement, setting expected levels of performance based upon these rates, and determining the extent to which its students, on the average, exceeded or fell short of expectation. Essentially, this is the same problem as establishing a school's average level of achievement after controlling for its students' previous levels of achievement. The procedures involved regression analysis to compute prediction equations by grade level and skill area independent of school identification and then using these equations within schools to obtain mean gains over expectations. The individual student was the unit of analysis.

Advantages of this Approach. There are a number of advantages that can be enumerated for this approach. First, it controlled for systematic influences governing the student composition of schools. Since, as many studies have shown, test score performance is highly correlated with student background, demographic, and environmental factors, controlling on previous test score performance indirectly controls for these other variables as well. Second, it provided all schools an equal opportunity to demonstrate success. Schools need only focus upon accelerating their own students' rates of achievement; they need not compete with each other in terms of absolute achievement levels. Thus, schools derived no particular advantage by being composed of higher- or lower-ability students. Third, since all schools are allocated resources on the basis of specific formulae (schools having similar needs are provided similar resources) the procedure was sensitive to differences in the way resources were managed. Finally, the approach was in consonance with many practitioners' views of what constituted an effective school.

Limitations. The major limitation of this approach was that it focused on standardized test performance. Certainly, learning and achievement occur in areas not measured by these tests, and in areas where the tests are less sensitive. While the equations developed through this approach proved to be very efficient and had a great deal of face validity, limiting the outcomes to standardized test results did not present a complete picture of the legitimate products of schooling. Standardized tests have been faulted in the literature for being insensitive to curricular content, focusing too much on recall of information (Frederickson, 1984), being biased in favor of more advantaged students (Guskey & Kifer, 1990), and for not measuring "real" achievement (Archibald & Newman, 1988). Kreft (1987) has argued that, since standardized tests are designed to distinguish between students and not between contexts, new tools are needed to yield more valid assessments of school outcomes. Thus, a design for effectiveness indices should contain an array of outcome measures in addition to standardized test results. Nevertheless, it must be realized that much of the recent public attention directed at the status of achievement in the nation's schools is really directed at the results of standardized testing programs. In the current educational environment, no matter what schools do to affect other outcomes, their efforts will be recognized largely to the extent that they affect standardized test scores. This was more true in 1984 than it is today.

A second limitation of this approach is not really a limitation, but rather a misperception of the method. The equations used three years of student achievement history in predicting student outcomes. The individual student growth curves carried with them important information about student background variables. It was demonstrated empirically that there was consistently no correlation between background variables such as ethnicity, free or reduced lunch, and gender, and school rankings by the equations. Nonetheless, the concern among practitioners that these variables were not accounted for in the equations continued, in spite of evidence to the contrary.

Finally, the issue of data aggregation is important. In the early school effectiveness studies, a single effectiveness index was derived. This necessarily involved several levels of data aggregation. Thus, within schools student test scores were aggregated by subtest (reading, math, or language) within grade level to form component subject area school effectiveness indices. These were then aggregated across subtests to form grade level school effectiveness indices. Finally, the grade level school effectiveness indices were aggregated to form a single, schoolwide school effectiveness

index. In examining intermediate results it was apparent that these steps often masked important effects among the components comprising the highest level of aggregation. For instance, in several instances a school's high (or low) rank could be traced to a particularly outstanding (positive or negative) effect at a single grade level in one subject area. Similar findings have been reported elsewhere (Abalos, Jolley, & Johnson, 1985; Helmstadter & Walton, 1985; Mandeville, 1988; Mandeville & Anderson, 1987) and cast doubt on the feasibility of aggregating school effectiveness indices over grade levels and subject areas without conscious thought as to the relative importance of each outcome variable.

Methodology. The first phase of this study involved computing 36 prediction equations, one for each possible combination of tests, subtests, and grade levels. In each equation the criterion was the spring '83 subtest score and the predictors (one at grade 2; two elsewhere) were the previous years' subtest scores. The accuracy of prediction was assessed by the standard errors of estimate, the indices of forecasting efficiency, and the multiple coefficients of determination (R^2). R^2's varied from a low of .28 for second grade mathematics to a high of .87 for eighth grade language. Most R^2's were .70 and above. Prediction generally improved with increasing grade level suggesting that, at higher grade levels, schools have a lesser opportunity to show a differential effect on student achievement as measured by standardized test scores. The use of three years of historical data on each student did not significantly improve prediction but did significantly reduce degrees of freedom in the equations.

The purpose of the next phase of the study, prediction and estimation, was to compute an aggregate measure of each school's actual performance with respect to its expected performance, the aggregation to be taken over students, grade levels, and tests. This phase involved several steps.

First, expected scores and differences between expected scores and actual scores for students within schools were computed. This was done separately by subtest within grade level and then aggregated over subtests and grade levels within schools.

For computing the individual difference scores, let:

Y_{sgi} be the Spring, 1983 grade equivalent score for individual i at grade level g on subtests, and

X_{1sgi}, X_{2sgi} be the Spring, 1981 and Spring, 1982 subtest scores for the same individual.

Then the expected 1983 score on s for each individual was given by the linear function:

$\hat{Y}_{sgi} = f_t (X_{1sgi}, X_{2sgi})$ where the function, f_t, was the prediction equation corresponding to the appropriate grade level, subtest, and battery (t) classification. The equations appeared as follows:

$$\hat{Y}_{sgin} = b_o + b_1 X_{1sgin-1} + b_2 X_{2sgin-2}.$$

Where:

$\hat{Y}_{sgin}$ = predicted outcome variable in year n for individual i at grade level g on subtest s.

b_0 = the constant

b_1 = the beta weight for year n-1

b_2 = the beta weight for year n-2

Individual difference scores were then computed as

$$dsgi = Y_{sgi} - \hat{Y}_{sgi}$$

The d_{sgi} eventually were to be aggregated over subtests and grade levels within schools. However, within a group to be aggregated, both the reliability and the scale of the individual d_{sgi} varied depending upon the particular function used to compute $\hat{Y}_{sgi}$ as well as the individual's location in the domain of predictors. Expectations at higher grade levels were generally more reliable than expectations at lower grade levels and the expectations for individuals close to the centroid of the predictors were more reliable than those for individuals further away.

To correct for differences in scale and reliability, the individual d_{sgi} were standardized by dividing by their individual standard errors of prediction to yield individually standardized scores.

After d_{sgi} were computed on each subtest for each individual in every school, they were then aggregated across subtests and grade levels within each school. A lower bound statistic was then computed by subtracting one standard error of the mean from the mean individually standardized residual score. This lower bound statistic, LB_d, was then used to rank the schools. The lower bound statistic allowed probability statements to be made regarding the placement of individual schools within the ranked distributions. Ranking was done separately within K-3, 4-6, K-6, 7-8 and 9-12 grade configurations.

In these early studies, the authors did not directly address the classic assumptions of the linear model (linearity, normality, homoscedasticity, independence). Instead, the equations were empirically validated through a series of studies that examined the extent of bias in the results toward schools with differing student characteristics. Correlations between school rank and student enrollment, percent white students, percent black students, percent Hispanic students, and percent students on free or reduced lunch programs were all non-significant. In addition, the ranking statistic was uncorrelated with schools' mean achievement for the previous year, a crucial result for establishing the fairness of the procedures.

These procedures were employed during the 1984-85 school year by the Dallas schools in a program to recognize and reward outstanding schools. Under this program, teachers in schools ranked in the top quarter of each grade-level category of schools each received a stipend of up to $1500. Other employees in those schools also received stipends, the amount determined by position and responsibility.

Effective Schools Methodology - 1992

School Effects. The basic rationale of the 1992 study is similar to that used in earlier studies. The school effectiveness methodology defined a school's effectiveness as being associated with exceptional measured performance above or below that which would be expected across the entire district. When a school's population of students departs markedly from the more general trend of similar students throughout the district,

this departure is attributed to school effect. The problem of measuring a school's effect, then, becomes one of establishing the student levels of accomplishment on the various important outcome variables, setting levels of performance based on these expectations, and determining the extent to which its students, on the average, exceed or fall short of expectation. The procedures involve regression analysis to compute prediction equations by grade level or by school for each outcome variable independent of school identification and then using these equations within schools to obtain mean gains over expectations. A major feature of this approach also involves assigning relative weights to each of the outcomes. Once weighted levels of performance have been determined, the methodology provides an indicator of how well a school performs relative to other schools throughout the district. Once again, as in the earlier studies, the student is the unit of analysis.

The difference between the logic of this approach and the one used in 1984 is subtle but substantive. In 1984 individual growth curves were developed for each student based on two years of historical test data. Equations were established at the student level with districtwide data, then the individual student residuals were applied to the students' home schools and aggregated. In the 1992 studies, as many predictor variables as was efficient were used to predict individual student achievement and other outcomes. Equations were again established at the student level with districtwide data, then the individual student residuals were applied to the students' home schools and aggregated. While individual student growth curves were used in 1984, the 1992 studies computed relationships at a systemwide level and then applied them to the schools. In 1984, an effective school was defined as a school that had more than half of its students exceeding prediction based on their individual growth curves. In 1992, an effective school was defined as a school that had more than half of its students exceeding prediction based on the patterns of other like students throughout the district.

There are a number of differences between the equations developed for the 1984 studies and the ones developed for 1992. First, and probably most important, the number and nature of outcomes were greatly expanded in 1992. While the 1984 studies used only standardized achievement tests, the 1992 studies added 143 separate course related criterion-referenced tests, student promotion and graduation rates, student attendance rates, and percentage of students taking and average scores on the *Scholastic Aptitude Tests (SAT).*

Changes in the Model. In addition to the expanded number of outcome variables used in the equations, a number of other changes were made in the model. First, an attempt was made to initially meet the assumptions of the linear model rather than to empirically validate the equations after the fact. The major assumptions that had to be met in order for the equations to produce adequate prediction were linearity and homoscedasticity. (Normality and independence of observations are only important if one is attempting to infer attributes from a sample to a population). Nonlinearity would be unacceptable because the equations would not fit the data. Non-homoscedasticity would be unacceptable because a student's prior position in the predictor distribution would bias the range of the residuals depending on the standard deviation of the residuals at a given point in the distribution. Linearity was routinely checked with each equation and achieved. Homoscedasticity was achieved by dividing each distribution into 128 arrays and normalizing each array around the regression line. This technique prevented similar scores from different points in the distribution from carrying more weight than

like scores at other points in the distribution. In addition, the regression lines representing all combinations of background variables were graphed to study patterns of bias in prediction at different points in the distribution. No consistent patterns were found.

The problem of perceived bias toward schools serving specific populations was addressed by using a two-stage process. In the first stage, each predictor and outcome variable was regressed on the set of important background variables and their first order interactions. Important background variables were ethnicity, gender, limited English proficiency status, and free or reduced lunch status. Residuals from these regressions then became the predictor and criterion variables for the next level of prediction. This demonstrably addressed practitioners' concerns about the impact of background variables on outcomes.

Once the initial residuals were obtained, a stepwise regression approach was used. As many predictors as was necessary to attain adequate prediction were included in the equations. As a result of using residuals that accounted for all of the student background information and of expanding the number of predictor variables, satisfactory prediction was attained without having to go back more than one year. (Most multiple r's were above .620 and 40% exceeded .700.) This maintained the degrees of freedom associated with the equations.

Effectiveness indices were produced for each outcome variable for each grade level by each of the combinations of background variables (gender by ethnicity, gender by LEP status, gender by economic status, etc.). All indices were standardized to a mean of 50 and a standard deviation of 10. This made them easily interpretable and provided a powerful diagnostic tool to determine patterns of service to various student groups.

Two additional aspects of the regression model were examined: the correlation of residuals with predicted values and the equality of outcome and predictor residual scores by the fairness variables. Correlations of residuals and predicted values would be unacceptable because it would imply that a student's prior position in the predictor distribution would bias the student's effectiveness score, the residual. Unequal residuals for any fairness variable group would suggest biased results within that group. Any of these problems would bias the effectiveness index for a school with a preponderance of such students from a given location in the distribution, a situation that the entire process was created to remedy.

Correlations of standardized residuals with predicted values were all near enough to zero to be acceptable. The largest was -.042. Again, statistical significance was not examined in these tests of aspects of the model because of the large numbers of degrees of freedom. Whether or not a result was statistically significant was irrelevant. Practical effect on the results was the only relevant criterion. The equality of residual means by fairness variable grouping was examined for each predictor and outcome variable. Examination of the means by fairness variable before and after the first regression stage showed that large pre-existing differences in group means were equalized by the procedures described earlier. Only small, non-practical differences between group means existed after the first regression stage.

Methodology. In the first regression and prediction phase, each predictor and outcome variable was regressed on a combined ethnicity/language proficiency variable, gender, and free/reduced lunch status and the first order interactions of these variables.

For computing the individual difference scores (residuals) in the first phase, let:

Y_{mgi} = the outcome variable of interest for each individual i at grade level g on measure m

X_{1mgi} = black status (1 if black, 0 otherwise)

X_{2mgi} = Hispanic English Proficient status (1 if English Proficient Hispanic, 0 otherwise)

X_{3mgi} = Hispanic LEP status (1 if Limited English Proficient Hispanic, 0 otherwise)

X_{4mgi} = gender status (1 if male, 0 otherwise)

X_{5mgi} = free or reduced lunch status (1 if subsidized lunch, 0 otherwise)

Then the expected 1992 level on each measure was given by the linear function

$$\hat{Y}_{mgi} = f_t (X_{1mgi}, X_{2mgi}, X_{3mgi}, X_{4mgi}, X_{5mgi}, X_{1mgi}X_{4mgi}, X_{2mgi}X_{4mgi}, X_{3mgi}X_{4mgi}, X_{1mgi}X_{5mgi}, X_{2mgi}X_{5mgi}, X_{3mgi}X_{5mgi}, X_{4mgi}X_{5mgi})$$

Where the function f_t, was the prediction equation corresponding to the appropriate grade level, measure, and outcome (t) classification. The equations appeared as follows:

$$\hat{Y}_{mgi} = b_o + b_1X_{1mgi} + b_2X_{2mgi} + b_3X_{3mgi} + b_4X_{4mgi} + b_5X_{5mgi} + b_6X_{1mgi}X_{4mgi} + b_7X_{2mgi} + b_8X_{3mgi}X_{4mgi} + b_9X_{1mgi}X_{5mgi} + b_{10}X_{2mgi}X_{5mgi} + b_{11}X_{3mgi}X_{5mgi} + b_{12}X_{4mgi}X_{5mgi}$$

Where:

$\hat{Y}_{mgi}$ = predicted outcome or predictor variable
b_o = the constant
b_1 = black student status
b_2 = Hispanic English proficient status (HEP)
b_3 = Hispanic Limited-English proficient status (HLEP)
b_4 = gender status
b_5 = free/reduced lunch status (FRL)
b_6 = black/gender interaction status
b_7 = HEP/gender interaction status
b_8 = HLEP/gender interaction status
b_9 = black/FRL interaction status

b_{10} = HEP/FRL interaction status

b_{11} = HLEP/FRL interaction status

b_{12} = gender/FRL interaction status

It should be noted that since a stepwise regression approach was used, the status variables were entered in different orders depending on the relationships between and among variables for each outcome and each predictor variable at each grade level. In not all cases were all background variables and interactions significant. It should be further noted that other students (non-black, non-HLEP, non-HEP) were not explicitly included in the first-stage equations since they formed the referent against the other ethnic/language students to avoid singularity of the regression design matrix.

From these equations, a predicted score, $\hat{Y}_{mgi}$, was computed for every student on each outcome variable at each grade. Also, predicted scores $\hat{X}_{1mgi}$, $\hat{X}_{2mgi}$, ... $\hat{X}_{nmgi}$, were computed for the n predictor variables for each student and grade.

Individual difference scores were then computed as

$$d_{mgi} = Y_{mgi} - \hat{Y}_{mgi} \text{ for criterion scores and}$$

$$X_{mgi} - \hat{X}_{mgi} \text{ for predictor scores.}$$

The multiple correlations between fairness variables and student-level outcome and predictor variables ranged from a low of .070 for student attendance at grade 12 to a high of .488 for *NAPT* Reading at grade 10. Coefficients were generally above .425 for *ITBS/NAPT* Reading, Vocabulary and Language tests and above .35 for the Mathematics test. Number of students ranged from a low of 3,705 for the correlation with the criterion-referenced tests at grade 12 to a high of 10,452 with student attendance at grade 1.

In the second phase of the study, the outcome residuals (d_{mgi}) computed during Phase 1 were regressed on the residualized predictor variables. Thus residuals of the predictor variables were used to predict residuals of the outcome variables. Once again, a stepwise regression approach was used. In all cases, the prior level of the outcome variable was the most significant predictor of the outcome variable. The equations appeared as follows:

$$\hat{Y}_{rmgi} = b_o + b_1 \hat{X}_{r1mgi} \ldots + b_n \hat{X}_{rnmgi}$$

Where:

$\hat{Y}_{rmgi}$ = predicted outcome variables (residuals) for each individual i at grade level g on measure m. All $\hat{Y}_r$'s and $\hat{X}_r$'s in this phase were residuals computed in the First Regression and Prediction Phase.

b_o = constant

b_1 = first predictor variable into the equation (residuals)

b_n = last predictor variable into the equation (residuals)

In the prediction of residual *outcome* variables from residual *predictor* variables, multiple r's ranged from .367 for grade 1 language to .818 for grade 8 language (with the exception of grade 1 attendance where kindergarten attendance was not available for a predictor and the multiple r was .156.) Most multiple r's were above .620 and over 40 percent exceeded .700. Degrees of freedom ranged from 3,596 for grade 11 language to 9,479 for grade 3 attendance. Most elementary degrees of freedom exceeded 7,000, most middle school exceeded 6,200, and most high school exceeded 4,000.

The next step of the second stage was to compute residuals (d_{rmgi}) from the regression equations. Raw residuals were computed using the following equation:

$$d_{rmgi} = Y_{rmgi} - \hat{Y}_{rmgi}$$

After raw residuals were computed, the predictor space was divided into 128 equal intervals. For each interval p, the mean and standard deviation of the raw residuals was computed. Each raw residual was then standardized by subtracting the mean for the interval and dividing by the standard deviation of the interval. Expressed as an equation, the standardized residuals were:

$$d_{prmgi} = (d_{prmgi} - \bar{d}_{prmg})/sd_{prmg}$$

Where:

d_{prmgi} = the standardized residual for individual i in interval p at grade g on measure m

$\bar{d}_{prmg}$ = the mean residual in interval p at grade g on measure m

sd_{prmgi} = the standard deviation of the residuals in interval p at grade g on measure m

The last part of the second phase was to determine the residuals for the school level variables. These were promotion rate at the elementary level and graduation rate and an *SAT* achievement/participation variable at the high school level. For this part of the regression analysis, the critical issue was degrees of freedom. Since there were a relatively large number of measures available and the school was the unit of analysis, the number of schools was relatively small and the models were easily overspecified if too many variables were entered. Further, it was impossible to tell where overspecification became a significant factor after the first variable was entered in the equation. Therefore, only one predictor variable was used for each school-level equation. School residuals were found by subtracting the predicted value of each school level variable from the actual value using simple regression equations.

The school ranking phase began at this point. The student-level part of the process will be described first. To simplify the notation from this point forward the following substitution will be made:

r_{smgi} = the standardized residual for individual i at grade g in school s on measure m

Mean residuals for each school on each measure at each grade were computed. In order to obtain rankings reflecting weighting of variables by sample size, the mean residuals were standardized by subtracting the district mean residual (approximately 0 after the standardization within intervals), dividing by the district standard deviation of

residuals (approximately 1), and multiplying by the square root of the n for each school on measure m at grade g. Expressing this as an equation:

$$M_{smg} = ((r_{smg} - r_{mg})/sd_{mg}) * \sqrt{n_{smg}}$$

Where:

M_{smg} = the mean deviation of the mean residual for school s at grade g on measure m (expressed in standard errors of the mean)

r_{smg} = the mean standardized residual at grade g in school s on measure m

r_{mg} = the district mean standardized residual at grade g on measure m

sd_{mg} = the district standard deviation of the standardized residuals at grade g on measure m

n_{smg} = the number of students at grade g in school s on measure m

The mean deviations M_{smg} were then ranked to produce the ranking of the schools on measure m at grade g. To combine mean deviations across measures to determine each school's effectiveness index the mean and standard deviation of the M_{smg} were computed for each m and g and the distribution of mean deviations standardized and expressed as a T-score. As an equation this is expressed:

$$T_{smg} = ((M_{smg} - M_{mg})/sd_{Mmg}) * 10 + 50$$

Where:

T_{smg} = the standardized mean deviation for school s at grade g on measure m (expressed as a T-score)

M_{smg} = the mean deviation at grade g in school s on measure m

M_{mg} = the district mean of mean deviations at grade g on measure m

sd_{Mmg} = the district standard deviation of mean deviations at grade g on measure m

Recall that school-level variables were ranked on simple deviations from a one-predictor regression. These deviations were expressed as T-scores using a process identical to the previous one for the residuals from the single variable regression.

To obtain the school effectiveness index, these standardized mean deviations were multiplied by the weight for the measure at the given grade (or, for school-level variables for the given school), summed and divided by the sum of the weights to obtain the school index. This is expressed in the following equation:

$$T_s = (\Sigma\Sigma_{mg} T_{smg} W_{mg} + \Sigma_S T_{Ss} W_S)/(\Sigma\Sigma_{mg} W_{mg} + \Sigma_S W_S)$$

Where:

T_s = the effectiveness index for school s

T_{smg} = the standardized mean deviation at grade g in school s on measure m

W_{mg} = the weight for measure m at grade g

T_{Ss} = the standardized mean deviation for school-wide variable S in schools

W_S = the weight for school level variable S

This procedure produced an easily interpretable T statistic for each school.

Since accountability indices without information for diagnosis and improvement are of limited utility, the system of equations which generates the school effectiveness indices also generates a great deal of information designed to help the schools diagnose areas of weaknesses. Schools receive rankings, also expressed in T scores, on each variable by grade as well as corollary output of the contributions of each student subgroup to the rankings by measure and grade. These outputs are used at the campus level to obtain information about campus effectiveness and needed areas of improvement. When used in conjunction with skills analyses, these statistics become powerful diagnostic tools.

Results

Face Validity

Table 1 displays some of the demographic characteristics of the top 20% of schools, as defined by the 1992 methodology. The reader will note that effective schools, as defined by this methodology, come in all sizes and shapes. District statistics at the particular grade levels are also presented to provide the reader with a framework for interpretation of the information.

At the K-6 level, the most effective schools tended to have smaller enrollments than the average enrollment of district elementary schools. Enrollments ranged from a low of 193 to a high of 860. Ethnicities ranged from a high of 99.7% black, 90.4% Hispanic, and 64.5% white to a low of 3.5% black, .3% Hispanic, and 0% white. Most deprivation indices were above the district average of 69, ranging as high as 92, while the percentage of limited English proficient students ranged from a high of 57.9% to 0. In short, whether or not a school was ranked among the most effective could not be predicted from the demographics of the students that it served.

Table 1: Demographic Characteristics Of The Top Twenty Percent Of Effective Schools

Rank	G	E	%W	%B	%H	%DEP	%LEP	SR
1	K-3	555	0.4	98.9	0.7	87	0.1	20
2	K-6	238	4.5	15.3	79.7	84	57.9	94
3	K-3	447	2.5	80.1	17.1	84	8.3	26
4	K-3	194	0	98.0	1.5	77	1.5	58
5	K-3	573	0.9	57.4	41.7	80	36.5	77
6	K-6	529	0.2	96.4	3.5	92	2.6	39
7	4-6	193	0.5	85.3	12.7	75	5.7	60
8	4-6	336	1.5	64.0	34.2	87	17.6	83
9	K-6	518	64.5	18.5	10.2	20	2.7	11
10	K-6	462	54.4	16.2	28.4	33	3.5	4
11	4-6	398	0.8	75.5	23.5	71	15.3	88
12	K-6	539	51.9	11.0	31.4	40	15.6	8
13	K-6	656	50.0	27.5	21.7	36	13.4	9
14	K-6	830	37.0	37.8	23.4	51	13.1	26
15	K-6	776	0	99.7	0.3	75	0	50
16	K-3	214	0.5	87.3	12.2	64	8.4	107
17	K-6	630	63.3	7.8	26.0	27	15.3	12
18	K-6	680	0.2	99.2	0.6	51	0	32
19	K-6	569	0.2	99.0	0.9	85	0	30
20	K-6	741	58.6	11.0	20.8	36	12.0	17
21	K-6	860	0.2	88.8	9.6	93	5.0	71
22	K-6	702	4.4	3.5	90.4	81	52.1	86
23	K-6	697	39.0	27.5	29.8	47	22.7	17
24	K-6	571	3.0	20.5	76.1	92	53.9	91
25	K-6	483	55.4	11.3	30.9	18	3.5	3
26	K-6	382	41.1	37.1	19.5	21	0	39
27	K-6	331	0	99.7	0.3	64	0	62
District	K-6	592	16.1	43.7	38.2	69	23.2	

	G	E	%W	%B	%H	%DEP	%LEP	SR
1	7-8	693	13.6	30.8	52.4	85	31.2	10
2	7-8	668	30.3	37.6	29.7	45	15.4	6
3	7-8	888	18.2	16.3	63.2	64	29.9	9
4	7-8*	367	24.4	50.4	22.4	38	0	2
5	7-8	863	7.0	75.4	15.7	30	1.3	5
District	7-8	703	15.2	48.6	34.3	55	13.2	

	G	E	%W	%B	%H	%DEP	%LEP	SR
1	9-12	1129	0.4	97.9	1.7	32	0	16
2	9-12	1004	36.7	37.7	23.3	24	13.0	4
3	9-12*	3567	18.5	43.7	32.9	22	4.8	5
4	9-12*	129	46.9	31.5	17.7	9	0	1
5	9-12*	623	48.4	34.3	15.6	9	0	2
6	9-12*	644	10.2	60.0	25.5	27	1.7	8
District	9-12	1093	15.8	49.8	31.5	28	10.1	

The last column in Table 1 (SR) depicts the school rank on the percent of students passing all subtests of the *Texas Assessment of Academic Skills* (*TAAS*), the most recently reported state test. The top twenty-seven K-6 schools in the district on the effectiveness indices had composite ranks between 3 and 107 when ranked based on absolute achievement levels. It should be noted that the six schools that ranked in the top fifteen in the district on the *TAAS*, when no known non-school sources of variation were accounted for, were at least 50% white and had no deprivation index above 40.

At the 7-8 level, the most effective schools had enrollments varying from 367 to 888 and were from 7.0% to 30.3% white, 16.3 to 75.4% black, and 15.7% to 63.2% Hispanic. Deprivation indices varied from a low of 30 to a high of 85 and percent limited English proficient varied from 0 to 31.2%. Ranks on the effectiveness indices and the *TAAS* were closer at this level than at K-6. Part of the reason for this is that district middle schools do not differ as much demographically as do district elementary schools.

At the high school level, magnet schools dominated the rankings. Four of the top six schools were magnet schools. This finding was predictable since at this level magnet schools spend about twice as much per student as do comprehensive high schools. Notice, however, that the most effective high school in the district, a school that is 97.9% black, and had a deprivation index of 32, was only ranked 16th out of 28 high schools on the *TAAS*.

The schools identified by this process fit the perceptions of most practitioners of what constitutes effective schools. Thus, the process provided results that had a great deal of face validity.

Consistency of Results

To examine consistency of results across years, 1983-85 effectiveness data were used. Despite the fact that the old effectiveness indices were based entirely on individual student growth curves on standardized achievement tests, and it was hypothesized that top schools could not maintain their standing because each time they finished high their individual student growth curves were accelerated, the results were extremely consistent. Effective schools tended to remain effective while ineffective schools tended to remain ineffective. Visual scans of the results from the 1983-85 studies and the 1992 study, using two very different models, suggest some amazing consistency between those schools identified as effective or ineffective in 1992 versus those schools identified as effective or ineffective in 1983-85. These same scans of the 1983-85 data also suggested that principal changes had some impact on schools improving or failing to improve. Since numerous principal changes occurred between 1985 and 1992, empirical verification of similarities in rank was not pursued until it could be determined whether or not principal changes had a major impact on results.

To examine school effectiveness after a change in school principal, 1983-84 and 1984-85 data were used. Regression equations were computed for grade K-3 and grade K-6 school configurations for both school years. (Analyses at other school configurations were not possible because of insufficient data.) In each of the four regressions, the assumption was made that a simple linear model could describe the existing data structure, regardless of whether or not the schools had experienced changes in principals. This assumption was tested using tests of homogeneity of regression.

Specifically, the data structure for each of the four data sets was represented using two lines, one for schools with changes in principals and one for schools with no changes in principals. In effect, with the more complete model, no assumptions were made regarding equality of intercepts or slopes. The question was then whether the more complete model provided more predictive power than the reduced model (i.e., the single regression line model). In statistical terms, the question was whether the two intercepts *and* two slopes in the more complete model were equal. If they were unequal, differences in school effectiveness were noted after changes in principals; if they were equal, differences in school effectiveness were not noted after changes in principals.

Of the four tests of homogeneity of regression, three were significant. This suggests that differences in school effectiveness occur after changes in principals and makes any analysis of 1983-85 results versus 1992 results for consistency inappropriate because of the large number of principal changes that occurred during that time frame.

Adjustments to the Model

Aiken and West (1991) present the argument that when all predictor variables are entered into the same equation, the slope of the regression line on the criterion variable for each value of each of the predictor variables is uniquely computed. This eliminates the necessity of assuming parallel regression lines for the various predictor variables. In empirical checks of the parallelism of the regression lines in the current model, most were found to be parallel. One exception was the Hispanic Limited English Proficient group at the seventh grade level. Simulations are currently being run to determine the differences, if any, produced by the two models. (That is, the two stage process currently used versus the fully specified model.) If the two models produce dissimilar results, the Accountability Task Force must weigh the relative advantage of widespread political acceptance of the current model versus a possible slight advantage in accuracy of the fully specified model. It is expected that most results will be very similar.

Discussion

For 1992-93 the outcomes used for the effectiveness indices include a nationally normed standardized test (*ITBS*, Grades 1-2; *NAPT*, grades 3-11), a state-mandated criterion-referenced test which includes a writing sample (*TAAS*, grades 4, 8, 10), 143 separate course-related criterion-referenced tests (*ACP*, grades 7-12), student promotion rate (grades K-8), student graduation rate (grades 9-12), student attendance rate (grades K-12), and percentage of students taking the *SAT* and average scores on the *SAT* (grades 11 and 12).

For 1993-94, six more outcome variables will be added to the equations. These include dropout rate (grades 7-12), student enrollment in accelerated courses with associated *ACP* scores (grades 7-12), high school enrollment in advanced diploma plans (grades 9-12), post-graduate enrollment in college or business schools, percent tested and average scores on the Preliminary Scholastic Aptitude Tests (*PSAT*, grade 10), and post-graduate pursuits.

A significant change will occur in 1993-94 when about 25% of the *ACP*'s will have components which include performance tests.

Figure 1: Formative and Summative Indicators Available to DISD Schools

Indicators	Goal(s) Impacted	Date Available
* *TAAS* Results Disaggregated by Demographic Variables ** Demographic Variables include Gender, Ethnicity, Free or Reduced Lunch and LEP (E)	1, 2, 6	Fall (Grades 3, 7, 11)
	1, 2, 6	Spring (Grades 4, 8, 10) #
ITBS Results Disaggregated by Demographic Variables (E)	1, 2, 6	Spring (Grades K-2) #
**NAPT* Results Disaggregated by Demographic Variables (E)	1, 2, 6	Spring (Grades 3-11) #
* *ACP* Results Disaggregated by Teacher and Skills (E)	1, 2, 6, 7	Fall & Spring, (Grades 7-12)#
Reconstituted *TAAS, ITBS, NAPT,* Data (Class Lists and Skills Analyses) (E)	1, 2, 7	End of fourth week of school
Disaggregated Test Data by Program (Chapter 1, Reading Improvement, Bilingual, etc.) by School (E)	1, 2, 6	Fall
Portfolios of Student Work (C)	1, 2	Local Option
Performance Testing (C,E)	1, 2	Local Option/ACP #
Protocol Analysis (C)	1, 2	Local Option
Teacher Satisfaction with Teaching, Ranking of Importance of Educational Goals, Perception of Teacher Influence, and Degree of Seriousness of Schoolwide Issues (E)	1, 2, 4, 5, 7, 9	Winter (all grades)
Student-to-Volunteer Ratio (E,C)	3	Fall
Volunteer Hours-to-Students (E,C)	3	Fall
Parental Involvement Log (C)	3	Local Option
Parent School Expectations, Perception of School Climate, Needs, Involvement/Participation (E)	3, 4	Winter (all grades)
* Student and Teacher Attendance (E,C)	1, 2, 5	Each six-week period #
Teacher Grade Distributions (E,C)	1, 2, 6, 7, 9	Each six-week period
School Effectiveness Indices (E)	1, 2, 5, 9, 10	September
School Effectiveness Indices Disaggregated by Student Group (E)	1, 2, 5, 9, 10	September
Student Satisfaction with Learning, Academic Self-Concept, Family Emphasis on Education, Cohesion	1, 2, 4	Winter (grades 4-12)
Teacher Climate Survey (E) (8 scales)	4	Provided on request by EPS
Student Climate Survey, Grades 4-12 (E)	4	Provided on request by EPS
Sociograms of Informal Interaction (lunch, recess, faculty meetings, etc.) (C)	4	Local Option

Fig. 1/ Cont.

Figure 1/ Continued

School-Community Council Survey (E)	4	Fall and Spring
Assistance and Consultation Team (ACT) Surveys (global issues, case management, training on mental health principles) (E)	4	Fall and Spring
Measures of Mobility and Stability (E)	5	Fall
Percent Eligible Tested versus Average Daily Attendance (E)	5	Fall
Monitoring of Local School Accreditation Remedies (C)	6	Fall
Monitoring of Implementation of Local School Programs (C)	7	Local Option
Monitoring of Instructional Delivery (C)	1, 2, 4, 6, 7	Local Determination
Student Retention Rate (E)	7	Fall #
* Student Enrollment in Advanced Courses (E,C)	8	Fall, Spring #
* Student Enrollment in Honors (E,C)	8	Fall, Spring #
* Student Enrollment in Diploma Plans (E,C)	8	Fall, Spring #
Survey of Student Course Interest (Grades 7-12) (E)	8, 9	Provided on request by EPS
* Dropout Rate (E)	9	December #
* Graduation Rate (E)	9	Fall #
* *SAT/ACT* Participation Rates (E,C)	10	Fall #
* *SAT/ACT* Scores (E)	10	Fall #
TASP Results (E)	10	Provided by the State
Graduate Follow-Up (E)	19	Fall #
Student Post-Graduate Pursuits (E)	8, 9, 10	Fall #
PSAT Participation Rates (E)	10	Fall #
PSAT Scores (E)	10	Fall #

* An Academic Excellence Indicator

** *TAAS* is the *Texas Assessment of Academic Skills*, a State-administered criterion-referenced test. *ITBS* is the *Iowa Tests of Basic Skills*. *NAPT* is the *Norm-referenced Assessment Program for Texas*, a Texas version of the *ITBS*. ACPs *are 143 criterion-referenced course exams, grades 7-12.*

Performance items and detailed scoring protocols will be provided to the schools. Samples will be drawn and selected tests rescored. Performance test results will then be adjusted by their reliability and included as outcomes in the equations. Preliminary analyses of the results of performance tests suggest that they are much more difficult than the average norm-referenced test over the same material (Dryden, 1991).

The effectiveness indices are an important part of the three-tier system of accountability being implemented in the DISD (Webster & Edwards, 1993). Training modules for school staffs in keeping and scoring portfolios of student work, designing and scoring performance tests, conducting protocol analysis, developing teacher-made tests, interpreting and using data, and designing and conducting appropriate action research are being undertaken to broaden the information that is currently available on student performance. Great emphasis is being placed on providing data for diagnosis and improvement. Figure 1 displays the range of data currently available to the schools. Indicators that are collected centrally and provided to the schools are specified with an "E". Indicators that school staff are being trained to collect and maintain themselves are specified with a "C". State academic excellence indicators are asterisked while variables that are or will become outcome variables in the effectiveness indices are marked with a #.

The effectiveness indices include a wide-range of variables, chosen and weighted by an Accountability Task Force which represents a composite of district groups vitally interested in education, and adjusted for inputs that are not under the control of the schools. Schools derive no particular advantage by starting with high-scoring or low-scoring students of any particular ethnic or economic group, are only held accountable for the outcome levels of cohorts of their continuously enrolled students, and are held accountable for a broad array of important education outcomes in addition to standardized test scores. The effectiveness indices are designed to foster teamwork among school staffs within schools in that school staffs must, in order to achieve the necessary improvements, work together in a coordinated effort. For this reason, the program does not reward individual competition among teachers.

Dallas Independent School District schools and their staffs were eligible for cash awards for 1991-92 performance based on the school effectiveness methodology under the district's School Performance Improvement Awards Program. In September of 1992, 2.4 million dollars was distributed to effective schools and their employees. Half of the 2.4 million dollars was budgeted by the district, the other half came from the community. To qualify schools had to exceed prediction on the effectiveness indices, test 95% of their eligible students, and outgain the national norm group in at least fifty percent of their cohorts. Once a school was selected as an award winner, the school received $2000 for its activity fund, each member of its professional staff received $1000, and each member of its support staff received $500. This program is continuing in 1992-93. Appendix A contains the details of the program.

References

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

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park: Sage.

Archibald, D. A., & Newman, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary schools.* Reston, VA: National Association of Secondary School Principals.

Bano, S. M. (1985). *The logic of teacher incentives.* Washington, D.C: National Association of State Boards of Education.

Brookover, W. B., & Lezotte, L. W. (1979). *Changes in school characteristics coincident with changes in students achievement.* East Lansing: Michigan State University, College of Urban Development.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, California: Sage.

Clark, D. L., Lotto, L. S., & McCarthy, M. (1980). Factors associated with success in urban elementary schools. *Phi Delta Kappan, 61,* 467-470.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426-443.

Cohen, M. (1986). *Designing state education assessment systems.* Paper prepared for the Study Group on the National Assessment of Student Achievement.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analyses for the behavioral sciences* (1st edition). Hillsdale, NJ: Lawrence Erlbaum.

Cuban, L. (1987). Transforming the frog into a prince: Effective schools research, policy, and practice at the district level. In R. V. Carlson & E. R. Ducharne (Eds.), *School improvement - Theory and practice: A book of readings* (pp. 993-1029). Lanham, MD: University Press of America.

D'Amico, J. (1982). Each effective school may be one of a kind. *Educational Leadership, 40,* 61-62.

Darlington, R. B. (1990). *Regression and linear models.* New York: McGraw-Hill.

David, J. (1987). *Improving education with locally developed indicators.* New Brunswick, NJ: Center for Policy Research in Education, Eagleton Institute of Politics, Rutgers, the State University of New Jersey.

Dryden, M. (1991). *Evaluation of the 1990-91 South and West Dallas learning centers.* Dallas Independent School District, REIS91-017.

Edmonds, R. R. (1982). Programs of school improvement: An overview. *Educational Leadership, 40,* 4-11.

Eubanks, E., & Levine, D. U. (1983). A first look at effective schools projects in New York City and Milwaukee. *Phi Delta Kappan, 64* (10), 697-702.

Farrar, E., Neufeld, B., & Miles, M. B. (1983). *Review of effective schools programs in high schools: Implications for policy, practice, and research.* Final report of the National Commission on Excellence in Education. (ERIC No. ED228243)

Felter, M. (1989). *A method for the construction of differentiated school norms.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC No. ED312302)

Felter, M., & Carlson, D. (1985). Identification of exemplary schools on a large scale. In G. Austin & F. Garber (Eds.), *Research on exemplary schools*, (pp. 83-96). New York: Academic Press.

Firestone, W. A., & Herriott, R. E. (1982). Prescriptions for effective elementary schools don't fit secondary schools. *Educational Leadership, 40,* 51-53.

Frederickson, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 193-202.

Fuller, W. A. (1976). *Introduction to statistical time series.* New York: John Wiley.

Good, T. L., & Brophy, J. E. (1986). School effects. In M. C. Wittrock (Ed.), *Handbook of research on teaching,* (3rd ed.) (pp. 570-602). New York, NY: Macmillan.

Guskey, T. R., & Kifer, E. (1990). Ranking school districts on the basis of statewide results: Is it meaningful or misleading? *Educational Measurement: Issues and Practice, 9* (1), 11-16.

Helmstadter, G., & Walton, M. (1985). *The generalizability of residualized indexes of effective schooling.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Jaeger, R. (1992). Weak measurement serving presumptive policy. *Kappan, 74* (2), 88-128.

Kirst, M. (1986). New directions for state education data systems. *Education and Urban Society, 18* (2), 347-357.

Klitgaard, R. E., & Hall, G. R. (1973). *A statistical search for unusually effective schools.* Santa Monica, CA: Rand Corporation.

Kreft, G. G. (1987). *Models and methods for the measurement of school effects.* The Netherlands: University of Amsterdam.

MacKenzie, D. (1983). School effectiveness research: A synthesis and assessment. In P. Duttweiler (Ed.), *Education productivity and school effectiveness.* Austin, TX: Southwest Educational Development Laboratory.

Mandeville, G. (1988). School effectiveness indices revisited: Cross-year stability. *Journal of Educational Measurement, 25* (4), 349-356.

Mandeville, G., & Anderson, L. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of Educational Measurement, 25* (3), 203-216.

May, J. (1990). *Real world considerations in the development of an effective school incentive program,* ED 320 271.

Murnane, R. J. (1987). Improving education indicators and economic indicators: The same problems? *Educational Evaluation and Policy Analysis, 9,* 101-116.

Nelson, R. R. (1973). *Applied time series analysis.* San Francisco: Holden-Day.

Neufeld, B., Farrar, E., & Miles, M. (1983). *A review of effective schools research: The message for secondary schools.* Bloomington, IN: Phi Delta Kappa International.

Nicoll, R. (1989). *School accountability in the Mt. Diablo Unified School District.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Oakes, J. (1989). What educational indicators? The case for assessing the school context. *Educational Evaluation and Policy Analysis, 11,* 181-199.

Ohio Department of Education. (1981). *Effective Sshools Process.* Columbus, OH: Ohio Department of Education.

Olson, G. H., & Webster, W. J. (1990). *Proposed procedures for measuring school effects in the Dallas Independent School District.* Dallas, TX: Dallas Independent School District, REIS90-140.

Pollard, J. (1987). *Viewpoints from selected states on accreditation and cccountability.* Austin, TX: Southwest Educational Development Laboratory.

Purkey, S. C., & Smith, M.S. (1983). Effective schools - A review. *Elementary School Journal, 83,* 426-452.

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

Saka, T. (1989). *Indicators of school effectiveness: Which are the most valid and what impacts upon them?* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC No. ED 306277)

Schalock, M. D., Cowart, B., & Staebler, B. (1993). Teacher productivity revisited: Definition, theory, measurement and application. *Journal of Personnel Evaluation, 7* (2), 179-196.

Shavelson, R. J., McDonnell, L. M., Oakes, J., & Carey, N. (1987). *Indicator systems for monitoring mathematics and science education.* Santa Monica, CA: Rand Corporation.

Shoemaker, J., & Fraser, H. (1981). What principals can do: Some implications from studies of effective schooling. *Phi Delta Kappan, 63,* 178-182.

Stedman, L. C. (1985). A new look at the effective schools literature. *Urban Education, 20,* 295-326.

Stedman, L. C. (1987). It's time we changed the effective schools formula. *Phi Delta Kappan, 69* (3), 215-224.

Stedman, L. C. (1988). The effective schools formula still needs changing: A reply to brookover. *Phi Delta Kappan, 69* (6), 439-442.

Van de Geer, J. P. (1971). *Introduction to multivariate analysis for the social sciences.* San Francisco: W. H. Freeman.

Webster, W. J., & Olson, G. H. (1984). *An empirical approach to edentifying effective schools.* ERIC TM 860 721

Webster, W. J., & Olson, G. H. (1988). A quantitative procedure for the identification of effective schools. *Journal of Experimental Education, 56,* 213-219.

Webster, W. J., & Edwards, M. E. (1993). *An accountability system for school improvement.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, Georgia, April 12-16, 1993.

The Authors

WILLIAM J. WEBSTER is the Division Executive for Program Evaluation and Accountability Services for the Dallas Independent School District. He has been head of Dallas's nationally recognized Evaluation Division for twenty-three years. He holds a Ph.D from Michigan State University. His specialities include research administration, experimental design, educational measurement, and statistics.

ROBERT L. MENDRO is the Executive Director of Institutional Research for the Dallas Independent School District. His department is responsible for all systemwide academic data analysis for the district. He holds a Ph.D in Educational Psychology from the University of Colarado at Boulder. His specialities include analysis and presentation of educational data, experimental design, educational measurement, use of computers in educational research and data analysis, and statistics.

TED O. ALMAGUER is the Director of Organizational Research and Development for the Dallas Independent School District. He is responsible for working with schools to better utilize evaluation data. He holds a Ph.D from New Mexico State University. His specialities include analysis and presentation of educational data, experimental design, and mathematical statistics.

Appendix A

Dallas Independent School District

SCHOOL PERFORMANCE IMPROVEMENT AWARDS - 1992-93

One of the key ingredients of the Commission for Educational Excellence's recommendations was an awards plan for effective schools. For 1992-93, the Dallas Independent School District (DISD) has budgeted 1.5 million dollars for this system. The community will raise $900,000, making a total availability of 2.4 million dollars. The selection procedure for determining which schools win is completely objective and is designed to award schools and school staffs that show the most improvement on important outcomes of schooling.

1.0 Outcome Variables

For the 1992-93 school year, awards will be based on school performance on the following variables:

1.1 *Elementary Schools*

1.1.1 Student scores on the Reading, Language, Vocabulary (at grade levels tested), and Mathematics subtests of the *Norm-referenced Assessment Program for Texas (NAPT)*. The *NAPT* is the State replacement for the *Iowa Tests of Basic Skills (ITBS)* and *Tests of Achievement and Proficiency (TAP)*. Grades K-2 will be tested with the *ITBS*.

1.1.2 Promotion Rate (percentage of students promoted, summer school doesn't count).

1.1.3 Student Attendance

1.1.4 Student scores on the *Texas Assessment Of Academic Skills (TAAS)*, Grade 4, Reading, Writing, and Mathematics subtests.

1.1.5 A special test will be developed or purchased to test Spanish-dominant Limited English Proficient students. This test will be administered to Spanish-dominant Limited English Proficient (LEP) students who are ineligible to be tested with the *ITBS* or *NAPT*.

1.2 *Middle Schools*

1.2.1 Student scores on the Reading, language, and Mathematics subtests of the *NAPT*.

1.2.2 Promotion Rate (percentage of students promoted, summer school doesn't count).

1.2.3 Student Attendance

1.2.4 First and second semester student *(Assessment of Course Performance (ACP)* scores in mathematics (Math 7, Math 8, Math 7 Pre-Honors, Algebra I, Pre-Algebra 8); language arts (Language Arts 7, Language Arts 8, English 1); social studies (Texas History/Geography 7, U. S. History 8); and science (Life Science 7, Earth Science 8, Pre-Honors Earth Science).

1.2.5 First and second semester student *ACP* scores in ESOL I, II, and III.

Appendix A/ continued

1.2.6 First and second semester student *ACP* scores in Reading Improvement, Reading 7, and Reading 8.

1.2.7 Student scores on *Texas Assessment of Academic Skills (TAAS)*, Grade 8, Reading, Writing, and Mathematics subtests.

1.3 *High Schools*

1.3.1 Student scores on the Reading, Written Expression, and Mathematics subtests of the *NAPT*.

1.3.2 Student Graduation Rate (the percent of students who graduate by the Spring semester five years after they enrolled in the ninth grade).

1.3.3 Percentage of seniors who have ever taken the *Scholastic Aptitude Test (SAT) / American College Tests (ACT)*.

1.3.4 *SAT / ACT* Achievement (juniors and seniors, highest score).

1.3.5 Student Attendance

1.3.6 First and second semester student *ACP* scores in mathematics (Algebra, Algebra II, Algebra II Pre-Honors, Geometry Pre-Honors, and Geometry); language arts (English I, II, III, IV); social studies (U. S. Government, World History, World Geography, U. S. History); science (Physical Science, Applied Biology, Biology I, Chemistry, Physics); and World Languages (Spanish I, II, French I, II, German I).

1.3.7 Student scores on the *TAAS*, Grade 10, Reading, Writing, and Mathematics subtests.

1.3.8 Student *ACP* scores in ESOL I, II, III, IV.

1.3.9 First and second semester *ACP* scores in Reading Improvement.

2.0 Qualifying Schools

All schools that have the necessary outcome data and all students will be included in the outcome equations. However, in order to be eligible for a School Performance Improvement Award all schools must:

2.1 Test at least 95% of their eligible continuously enrolled students or increase their percent eligible continuously enrolled students tested by 3% over Spring, 1992. These statistics refer to percent tested on the *NAPT*. Students at the *School Community Guidance Center (SCGC)* will be tested and attributed to their home schools.

2.2 Test at least their percent average daily attendance for grades 4, 8, or 10 on the *TAAS*.

2.3 Exceed the national norm group growth curves, or be above the national norm group, in at least 50% of school cohorts on the *NAPT*.

If a school does not meet each of the aforementioned criteria, it will not be eligible for a School Performance Improvement Award.

Appendix A/ continued

3.0 Establishing School Cohorts

Since the School Performance Improvement Award is based entirely on student outcomes (once a school has qualified) it is important to specify which students will be included in the various cohorts. Therefore:

3.1 *Establishing School Cohorts*

All students who:

3.1.1 are enrolled continuously in a specific school from the end of the first six weeks, and

3.1.2 have the necessary pre-observation data in the DISD and post-observation data for the 1991-92 school year in that specific school, and

3.1.3 are eligible for the testing program according to the DISD Systemwide Testing Policy (on the testing variables)

will be included in the cohort longitudinal analysis. Thus, in order to be included as a member of a given school's cohort, a student must be enrolled in that school by the end of the first six weeks, have the necessary pre-observation data, and be tested in that school in accordance with DISD policy through the systemwide testing program. Students who transfer out of a school and back into that school over a short period of time will be included in that school's cohort. Schools that, in the opinion of the Accountability Task Force, attempt to manipulate their continuously enrolled student population will be disqualified from the Awards Program.

4.0 Qualifying Staff for Awards

Once a school has been empirically selected for a School Performance Improvement Award, the school will receive $2000 to be spent in a manner, other than compensation, to be determined by the School Community Council (SCC) Committee in School Centered Education (SCE) schools or the Faculty/Staff Advisory Committee in non-SCE schools. Performance awards will also be distributed in the form of compensation to the staff of winning schools based on the following criteria.

4.1 Eligible Staff

4.1.1 Principals will be eligible to receive a stipend.

4.1.2 All campus personnel will be eligible to receive a stipend if they are full-time professional or support personnel who are assigned to a single campus and are evaluated by a local campus administrator.

4.1.3 Professional or support personnel who are assigned to more than one campus and evaluated by one or more campus administrator(s) will receive a pro rata share of the stipend. Proration will be based on the percentage of time assigned to one or more winning schools.

4.1.4 In circumstances where there are variable hours worked within an employee classification the employee will receive a pro rata share based on the percentage they work of the standard work day of their respective classification.

Appendix A/ continued

4.2 Successful Evaluation

Individuals must be evaluated "Meets Expectations" or above in order to participate in monetary awards.

4.3 Stipends

4.3.1 *Professional Staff*

Stipends will be paid to professional staff who are assigned to winning schools. The amount of the stipend will be determined by the considerations specified in Section 4.1 and by attendance during the contract year.

4.3.1.1 *Attendance*

Eligible professional staff who are present all contract days of the school year and meet requirements 4.1.1 or 4.1.2 will receive a stipend of $1,000. Professional staff who are not present all contract days will receive an award of one thousand dollars minus five dollars per day for every contract day absent. If professional staff are not full-time at a winning school, their share will be calculated in the manner specified in 4.1.3 or 4.1.4.

4.3.2 *Support Staff*

Stipends will be paid to support staff who are assigned to winning schools. The amount of the stipend will be determined by the considerations specified in Section 4.1 and by attendance during the contract year.

4.3.2.1 *Attendance*

Eligible support staff who are present all contract days of the school year and meet requirements 4.1.1 or 4.1.2 will receive a stipend of $500. Support staff who are not present all contract days will receive an award of five hundred dollars minus $2.50 per day for every contract day absent. If support staff are not full-time at a winning school, their share will be calculated in the manner specified in 4.1.3 or 4.1.4.

5.0 Number of Winning Schools

The number of winning schools will depend on the size of the schools that win. There will be approximately 1,850 winning professional and 800 winning support personnel. The determining factor will be the number of staff associated with winning schools that can be awarded stipends of up to $1,000 and $500 for professional and support personnel, respectively, within the available 2.4 million dollars. (If a large number of large schools win, fewer schools will be included in the awards. Conversely, if a large number of small schools win, more schools will be included in the awards.

Appendix A/ continued

6.0 Establishing Appropriate Comparisons

In order to allow all school configurations a reasonable chance of receiving a School Performance Improvement Award, District schools will be chosen according to the following categories:

6.1 Categories for Comparison

	Grade Level
6.1	PK-3
6.2	4-6 and Vanguards
6.3	PK-6 and Vanguards
6.4	7-8 and Academies
6.5	9-12 and Magnets

The amount of money available for each level will be determined by the percentage of school-based professional personnel employed at each level.

6.2 Magnets, Vanguards, and Academies

Magnets, Vanguards, and Academies will be treated as separate programs at the appropriate level if they have separate teaching and administrative staffs. Otherwise, they will be included with the appropriate school. The following academies and vanguards, located in the same building with a comprehensive school, will be treated as separate schools:

Holmes Academy
Spence Academy
Lanier Vanguard
Polk Vanguard

6.3 Schools Not Meeting Standard Criteria

Several schools have insufficient data on one or more critical variables included in the school effectiveness indices and therefore cannot be included in the Award Program. These schools are not included in the regular process due to the nature of the school or the student enrollment at the school. In either case, school effects cannot be computed using the procedures proposed for the school effectiveness indices. The schools which are not yet included in the process for 1992-93 are:

Health Special
E. D. Walker Special Education Center
Multiple Career Center
Alternative Academic Cooperative Center
Evening Schools (Skyline and Kimball)
Metropolitan Education Center
School Community Guidance Center
Letot Academy
Brashear
Quentin D. Corley Academy
Edison Work Activity Center
Science Magnet

Appendix A/ continued

Ad Ad Hoc Committee of the Accountability Task Force, chaired by Dr. Herman Saettler has been appointed to work with these schools in producing an appropriate plan for the 1992-93 school year.

6.4 Employees Not Meeting Standard Criteria

Classifications of employees who are, because of budgetary or supervisory criteria, excluded from participation in this program are invited to submit ideas and/or proposals that might achieve the same goals for their respective groups. These proposals should be submitted to Robby Collins, Executive Manager, Governmental/Internal Relations, 3700 Ross Avenue, Box 9. All proposals will be considered by the Accountability Task Force for possible implementation.

7.0 The Equations

The school effectiveness methodology defines a school's effectiveness as being associated with exceptional measured performance above or below that which would be expected across the entire District. When a school's population of students departs markedly from its own pre-established trend or from the more general trend of similar students throughout the District, this departure is attributed to school effect. The problem of measuring a school's effect, then, becomes one of establishing the student levels of accomplishment on the various important outcome variables, setting levels of performance based on these expectations, and determining the extent to which its students, on the average, exceed or fall short of expectation. The procedures involve regression analysis to compute prediction equations by grade level or by school for each outcome variable independent of school identification and then using these equations within schools to obtain mean gains over expectations. A major feature of this approach also involves assigning relative weights to each of the outcomes. Once weighted levels of performance have been determined, the methodology provides an indicator of how well a school performs relative to other schools throughout the District. Important characteristics of the methodology include:

7.1 Schools are only held accountable for the outcome levels of students who have been exposed to that school's instructional program. That is, schools are only held accountable for their continuously enrolled students.

7.2 The influence of important background variables of students, over which the schools have no control, are eliminated from the equations. That is, each predictor and outcome variable is regressed on the set of background variables (ethnicity, gender, limited English proficiency status, and free or reduced lunch status) and residuals from these regressions then become the predictor and criterion variables for the next level of prediction. This "levels the playing field" and addresses practitioners' concerns about the impact of background variables on outcomes. Other fairness variables including, but not limited to, size of school, overcrowding conditions, etc., will be examined for inclusion.

7.3 The outcome variables are weighted by the Accountability Task Force.

7.4 Schools derive no advantage by starting with high-scoring or low-scoring students. That is, the equations set individual expectations for each student based on that student's placement on the pretest(s) of interest. Lower scoring students have lower predicted scores. Higher scoring students have higher predicted scores.

Appendix A/ continued

7.5 Only one year of historical data are used. That is a stepwise regression approach is used on the residuals of multiple predictors so that in most cases satisfactory prediction is achieved without having to go back more than one year. This maintains the degrees of freedom associated with the equations since, in an urban district, each additional year of data used significantly reduces the degrees of freedom associated with the equations.

8.0 Outcome Variables for 1993-94

There are a number of outcome variables that, because of timeliness, are not included for 1992-93 but will be included for 1993-94. These include:

8.1 Dropout rate for middle and high schools. Because of the time that dropout rate becomes available, this will be a time-lag design with 1992-93 dropout rate being included in the 1993-94 equations.

8.2 Student enrollment in accelerated courses for middle and high schools as well as ACP scores in each.

8.3 High school student enrollment in Advanced Diploma Plans.

8.4 Post-graduate enrollment in college or business schools. This will also be a time-lag design with the graduate follow-up of the 1992-93 Class used in the 1993-94 equations.

8.5 Percent tested and *preliminary Scholastic Aptitude Test (PSAT)* achieve-ment.

8.6 Post-graduate career pursuits.

8.7 The weight for Graduation Rate for high schools will increase by one point per year.

9.0 Weights of Outcome Variables

For the 1992-93 school year, outcome variables will have the following weights:

Grade	1	2	3	4	5	6	7	8	9	10	11	12
ITBS												
Reading	3	3	*	*	*	*	*	*	*	*	*	*
Language	2	2	*	*	*	*	*	*	*	*	*	*
Vocabulary	2	2	*	*	*	*	*	*	*	*	*	*
Math	2	2	*	*	*	*	*	*	*	*	*	*
NAPT												
Reading	*	*	3	3	3	3	4	4	3	3	3	*
Language	*	*	2	2	2	2	2	2	3	3	3	*
Math	*	*	2	2	2	2	3	3	3	3	3	*
Promotion	**1 per school**						1		*	*	*	*

Appendix A/ continued

Weights of Outcome Variables, cont.

Grade	1	2	3	4	5	6	7	8	9	10	11	12
Attendance	1	1	1	1	1	1	1	1	1	1	1	1
TAAS												
Reading	*	*	*	3	*	*	*	3	*	4	*	*
Writing	*	*	*	3	*	*	*	3	*	4	*	*
Math	*	*	*	2	*	*	*	2	*	4	*	*
LEP Test	3	3	3	3	3	3	*	*	*	*	*	*
ACP												
Language Arts	*	*	*	*	*	*	2	2	2	2	2	2
Math	*	*	*	*	*	*	2	2	2	2	2	2
Social Studies	*	*	*	*	*	*	2	2	2	2	2	2
Science	*	*	*	*	*	*	2	2	2	2	2	2
ESOL	*	*	*	*	*	*	2	2	2	2	2	2
Reading	*	*	*	*	*	*	2	2	2	*	*	*
World Language	*	*	*	*	*	*	*	*	1	1	1	1
Graduation Rate	*	*	*	*	*	*	*	*	5 per school			
SAT/ACT % Tested	*	*	*	*	*	*	*	*	*	*	5	
SAT/ACT Score	*	*	*	*	*	*	*	*	*	*	4	