Measuring the Effects of Schooling: Expanded School Effectiveness Indices

William J. Webster, Robert L. Mendro, and Ted O. Admager
Dallas Independent School District

The need for instructional improvement in America’s schools has been thoroughly documented and publicized over the past few years. One of the responses to this need has been the effective schools movement. Researchers investigating effective schools have consistently identified five to seven factors correlated with improved school achievement (Good & Brophy, 1966; Purkey & Smith, 1983). These factors included a sense of mission (Brookover & Lexow, 1979; Ohio Department of Education, 1982; Venesky & Winfield, 1979); strong building leadership (Edmonds, 1982; Shoemaker & Fraser, 1981); high expectations for students and staff (Clark, Loco, & McCarthy, 1980; E-banks & Levine, 1983); frequent monitoring of student progress (Edmonds, 1982; Venesky & Winfield, 1979); a positive, orderly learning climate (Edmonds, 1982; Shoemaker & Fraser, 1981); sufficient opportunity for learning (Brookover, Beady, Flood, Schweitzer, & Wisentaker, 1979; Edmonds, 1982; MacKenzie, 1983); and parent/community involvement (Ohio Department of Education, 1982; 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 (Cuban, 1987; Good & Brophy, 1986; Purkey & Smith, 1983). More specifically, concerns posed by a number of investigators include the following:

1. The development of new techniques for evaluating school effectiveness has not kept pace with the increased interest (Feltus, 1989; Webster & Olson, 1988).

2. Most studies have focused upon narrow educational outcomes. Typically, academic achievement has been used as the only indicator of school effectiveness. Further, the only indicator was often a norm-referenced score for reading and/or mathematics (Rowan, Bossert, & Dwyer, 1983; Stedman, 1987; Stedman, 1988).

3. The research has been primarily limited to elementary schools in urban systems with large populations of disadvantaged youth (Clark, Loto, & McCarthy, 1980; Farrar, Seufeld, & Miles, 1983; Firestone & Herriott, 1982; Rowan, Bossert, & Dwyer, 1983).

2 Paper updated based on further simulations and empirical results to reflect the process used for 1991-92.
4. Most of the research attempting to associate school effects with student learning is correlational. Effective schools research cannot claim that any set of correlates cause a school to be instructionally effective. There are as yet no recipes to make effective schools (D'Amico, 1982; Neufeld, Farrar, & Miles, 1983; Purkey & Smith, 1983).

If research related to effective schools is to be advanced, new techniques for evaluating effective schools need to be developed (Saka, 1989). Inherent in this task are two complex issues: (a) a better definition of effectiveness and (b) development of a model within which it can be assessed.

**Improved Definition**

In attempts to provide a better definition of effectiveness and responding to the narrowly focused concern of earlier effective schools research, Murmane (1987), David (1987), and others have been proponents for developing an expanded number of outcome indicators. In addition, Oakes (1989), David (1987), and Cohen (1986) have argued the importance of incorporating input and process/context indicators as important aspects of better accountability mechanisms.

Possible input indicators often include school enrollment, socioeconomic/ethnic composition, proportion of limited English speaking children, enrollments in categorical programs, staff characteristics, and financial resources. Process indicators describe what is being taught, the way it is being taught, and include consensus on school goals, instructional leadership, opportunity to learn, school climate, staff development, and collegial interaction among teachers. Outcome indicators are usually related to capturing the results of school on students or providing information about other definitions of "good schooling," and may include students' academic performance, teacher and student attendance rates, dropout and completion rates, performance of students at the next level of schooling, parent and student satisfaction, percent completing advanced courses, college attendance, and individual school goals (David, 1987; Oakes, 1989; Olson & Webster, 1990; Pollard, 1987; Shavelson, McDonald, Oakes, & Carey, 1987).

**Improved Assessment Model**

The commonly utilized technique of comparing schools based on outcome measures can adversely affect those schools with less-than-ideal population demographics (Felter, 1989; Grobe, Sheehan, Sides, & Read, 1989). For example, students in one school may have more ability to learn than students in another school. The non-statistical technique of comparing schools with similar characteristics is one viable solution for cases involving a limited number of grouping characteristics (Nicoll, 1988; Felter, 1989).
However, to incorporate a large number of input, process, and outcome variables, an appropriate statistical method is multiple regression analysis (Barre 1976; Felter & Carlson, 1985; Kief, 1986; Kliggaard & Hall, 1973; MacKenzie, 1983; Sakr, 1989). As a simplified illustration, the mean score for an outcome measure such as achievement is predicted after considering such input variables as gender, ethnicity, and initial ability level. The difference between predicted and actual achievement, a residual or adjusted score, can be interpreted as a comparison with other statistically similar schools, and as the school's own effect on achievement.

Previous Studies

There are two previous series of studies that have direct bearing on the present studies. These studies were those conducted in Dallas in 1984-88 (Webster and Olson, 1988) and by the South Carolina Department of Education in 1985 to the present (May, 1990). Both used a form of regression analysis to establish expectations against which actual student performance was compared. While the South Carolina studies initially used cross-sectional data, the investigators rapidly concluded that the use of longitudinal cohort data was the appropriate course of action. The Dallas studies used longitudinal cohort data.

In the early Dallas 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 the student’s previous rate of achievement growth. In general it is reasonable to expect students to continue to achieve at a given rate. Thus, when a school’s population of students departs markedly from its own pre-established trend or from the more general trend of similarly achieving students throughout the district, this departure is 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 involve using 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 expectation.

These are a number of advantages that can be listed for this approach. First, it controls for systematic influences related to 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 gives 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 derive no particular advantage by being composed of higher- or lower-ability students. Third, since all schools are allocated resources on the basis of specific formulas (schools having similar needs are provided similar resources) the procedure is sensitive to differences in the way resources are managed. Finally, the approach is in consonance with many practitioners' views of what constitutes an effective school.

One limitation of this approach is 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 and Kifer, 1990), and for not measuring "real" achievement (Archie and Newman, 1988). Kreft (1988) 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.

A second limitation to this approach is not really a limitation, but rather a misperception of the method. The equations utilized 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 students' 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, Jolly, & Johnson, 1985; Helmsdale & Walto, 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.

The Present Study

The basic rationale of the present study is similar to that used in earlier studies. 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.

The major differences between the approach discussed in this paper and previous studies are:

1.0 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.

2.9 A stepwise regression approach is used on the residuals 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.

3.0 Outcome variables are weighted by an Accountability Task Force rather than all variables being given the same weight.

4.0 Once the background variables are removed, problems related to the homoscedasticity of the distributions are exacerbated. Thus, adjustments to the prediction equations must be made. Each distribution was divided into 128 arrays and each array was normalized.

5.0 A wider range of outcome and predictor variables are used in the equation. Ultimately to be included among the outcome variables are Iowa Test of Basic Skills results, Tests of Achievement and Proficiency results, Norm-referenced Assessment Program for Texas (state standardized test norm-equated to the ITBS/ITAP), Texas Assessment of Academic Skills results (State test), STEELS results (District curriculum-referenced tests in 143 different courses), student dropout indices, student graduation rates, student attendance rates, teacher attendance rates, percentage of students with advanced diploma plans, percentage of students taking the Scholastic Aptitude Tests and the American College Testing Program and average scores on each, percentage of graduates attending post-secondary education, exit rate from remedial programs, student course passing rate, and enrollment rate in advanced courses.

Cases and Variables:

Cases: Cases included in this study were all Dallas students grades 1-12 who had the appropriate data in Spring, 1992, and who met the following eligibility criteria:

(a) were enrolled continuously in a specific school from the end of the first six weeks,
(b) had the necessary pre-observation scores in the Dallas District and had post-observation scores for the 1991-92 school year in that specific school, and
(c) were eligible for the systemwide testing program according to DISD Testing Policy.

All analyses were by cohort.

Data. Outcome variables for the 1991-92 school year included:

1.0 Elementary School

3 A discussion of the Accountability Task Force is included in the discussion section of this paper.
1.1 Student scores on the Reading, Language, Vocabulary, and Mathematics subtests of the Iowa Tests of Basic Skills (ITBS), and Norm-Referenced Assessment Program for Texas (NAFT).

1.2 Promotion Rate (percentage of students promoted).

1.3 Student Attendance.

2.0 Middle Schools

2.1 Student scores on the Reading, Language, Vocabulary, and Mathematics subtests of the ITBS, and NAFT.

2.2 Student scores on the STEELS (District curriculum-referenced tests in 143 different subjects).

2.3 Student Attendance.

3.0 High Schools

3.1 Student scores on the Reading, Language, Vocabulary, and Mathematics subtests of the NAFT.

3.2 Student scores on the STEELS.

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

3.4 Percentage of graduating seniors scoring above the District mean on the SAT.

3.5 Student Attendance.

In this first year, data on other outcome variables to eventually be included in these studies were not available because of the stringent timelines involved in the development and implementation of the equations.

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. The equations were as follows:

\[ \hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8 + b_9X_9 + b_{10}X_{10} + b_{11}X_{11} + b_{12}X_{12} \]
Where:

\[ \hat{Y} = \text{predicted outcome or predictor variable} \]

- \( b_0 \) = constant
- \( b_1 \) = Black student status
- \( b_2 \) = Hispanic English proficient student status (HEP)
- \( b_3 \) = Hispanic Limited-English proficient student 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, \( \hat{Y} \), was computed for every student on each outcome variable at each grade. Also, predicted scores \( \hat{X}_i, i=1,2,..., N \) were computed for the \( N \) predictor variables for each student and grade. Then each student's predicted scores, \( \hat{Y} \) and \( \hat{X}_i \), were subtracted from the actual scores, \( Y \) and \( X_i \), respectively, to yield a set of individual student residuals on predictor and outcome variables. Thus, the first regression and prediction phase was
computed with residual scores for each student on each outcome and predictor variable at each grade level.

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. Numbers of students ranged from a low of 3,705 for the correlation with STEELS at grade 12 to a high of 10,452 with student attendance at grade 1. At the elementary level, most N’s were above 8,200, at the middle school level above 7,000, and at the high school level above 4,600.

**Second Regression and Prediction Phase**

In the second phase of the study, the outcome residuals 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:

$$Y_k = b_0 + b_1 X_1 + \ldots + b_n X_n$$

Where:

- $Y_k$ = predicted outcome variables (residuals). All $Y_k$'s and $X_k$'s in this phase were residuals computed in the First Regression and Prediction Phase.
- $b_0$ = constant
- $b_1$ = first predictor variable into the equation (residuals)
- $b_n$ = last predictor variable into the equation (residuals)

At this point, decisions had to be made regarding the ultimate utility of the results. Inclusion of multiple predictor variables, particularly more than one year of prior achievement, significantly reduced the degrees of freedom while not adding substantially to the efficiency of prediction. Because of the large numbers involved at every level, statistical significance was achieved in every instance and was not an issue. Practical significance was the criterion by which equations were chosen.

The issue to be determined was whether a small gain in predictability outweighed the reduction in the number of students on which prediction was based. In the real world, the
number of students involved in the equations is extremely important for philosophical reasons and practitioner acceptance. From a philosophical standpoint, school effectiveness must impact as many students as possible. From a practical standpoint, effectiveness procedures which excluded any significant number of students for any reason other than failure to be exposed to the school program would be unacceptable to the public and most educational practitioners. Thus, in most cases, the advantages of using two or more years of student achievement data to predict current achievement were sacrificed, at very little practical expense, in favor of retention of significantly larger numbers of students in the equations.

For example, consider the prediction of the residual ITBS Language subtest for 1991 in tests of the model from the previous year. Reducing the initial set of variables to the most significant, the best predictors were ITBS from 1989 and 1990 and student retention status. These gave an equation with a multiple \( r \) of .793 with 6,143 degrees of freedom across 27 schools. Reducing the set to the four ITBS subtests for 1990, the multiple \( r \) dropped to .790 and degrees of freedom increased to 6,693. Further reducing the predictors to ITBS Language and Mathematics for 1990 resulted in a slight increase in degrees of freedom to 6,713 with the multiple \( r \) remaining at .790. Finally, predicting the residual from ITBS Language for 1990, the multiple \( r \) dropped to .782 and degrees of freedom increased to 6,777. Thus, by going from 9 predictors to 1, the drop in multiple \( r \) was .011. In terms of multiple \( r^2 \), the drop was from .629 to .612 or approximately 1.7% of the explained variance, a negligible amount. The increase in degrees of freedom was 654, or 10.3%, with an average gain of more than 23 degrees of freedom per school at that grade. Most situations were similar to this example. In some prediction equations, however, the drop in multiple \( r \) exceeded acceptable levels (an empirical level of approximately a 5% reduction in multiple \( r^2 \) was used as a reference point with each set of equations being determined based on the corresponding increase in degrees of freedom). In these instances, fewer degrees of freedom were accepted in order to assure a more effective level of prediction.

Thus, in the actual determination of indices for 1991-92, the previous year's data were sufficient to provide excellent prediction without sacrificing large numbers of degrees of freedom. In most instances the full set of previous year predictors (for example, the ITBS subtests from 1991 used to predict the NAPRT subtests in 1992) were used in the equations with little reduction in degrees of freedom. In a few instances, several predictors were dropped because the loss of degrees of freedom was too high relative to the accuracy of prediction. For example, using residuals of 1990-91 NAPRT vocabulary, reading, language, and mathematics and 1990-91 STEELs to predict residual 1991-92 STEELs resulted in a multiple \( r \) of .821 with 5,884
Reducing predictors to STELLS alone resulted in a multiple $R$ of .814 with 6,380 degrees of freedom. Hence multiple $R^2$ dropped from .673 to .662 (less than 2%) while degrees of freedom increased from 5,984 to 6,380 (approximately 6.6%). In this case the reduced model was used.

In the prediction of residual outcome variables from residual predictor variables, multiple $R's$ ranged from .367 for grade 1 TEST language to .818 for grade 8 NAPT 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 NAPT 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.

Three aspects of the regression model were examined: the correlation of residuals with predicted values, homoscedasticity of residuals throughout the range of 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. Nonhomoscedasticity 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. 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 a result was statistically significant was irrelevant. Practical effect on the results was the only relevant criterion.). To examine homoscedasticity, the range of predictor values was broken into 128 equal arrays. The mean and standard deviation of the residuals for each part was computed and examined. There were small fluctuations in the mean and relatively larger fluctuations in the standard deviations. Thus, the examination of homoscedasticity showed that residuals were not homoscedastic or nearly enough so across the range of predictor values.

Across variables, no consistent pattern of mean values was determined. That is, means
Across variables, no consistent pattern of mean values was determined. That is, means fluctuated differently relative to predictor values with no recognizable system across variables or across grades within variable. Standard deviations tended to follow a pattern of larger values associated with lower values of the predictor variables. Although this pattern was not consistent with all variables. To control for the lack of homoscedasticity, residuals were standardized within each cell by subtracting the mean and dividing by the standard deviation. Based on data this approach is a viable solution to the problem.

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.

Rankings Phase

Once each student's standardized residual values on each variable were computed, the next step involved determining mean residuals for individual campuses. These mean residuals were then standardized within each distribution. To account for differences in school size, the means for each school were then standardized by computing the standard deviation of the mean for each school mean. Outcome variables were then weighted based on a weighting schema determined by the Accountability Task Force. Table 1 shows the weighting system devised by the Task Force. The weighting schema emphasizes standardized achievement at the elementary grades and places increasing emphasis on curriculum-referenced tests for measures of middle and high school achievement.

<table>
<thead>
<tr>
<th>Elementary Criteria</th>
<th>Weight</th>
<th>Middle School Criteria</th>
<th>Weight</th>
<th>High School Criteria</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAPK-Vocabulary</td>
<td>3/Grade</td>
<td>STEELS</td>
<td>2/Grade</td>
<td>NAPK-Reading</td>
<td>3/Grade</td>
</tr>
<tr>
<td>NAPT-Reading</td>
<td>2/Grade</td>
<td>NAPK-Vocabulary</td>
<td>4/Grade</td>
<td>NAPT-Reading</td>
<td>3/Grade</td>
</tr>
<tr>
<td>NAPT-Language</td>
<td>2/Grade</td>
<td>NAPK-Reading</td>
<td>4/Grade</td>
<td>NAPT-Language</td>
<td>3/Grade</td>
</tr>
<tr>
<td>NAPT-Mathematics</td>
<td>2/Grade</td>
<td>NAPK-Language</td>
<td>2/Grade</td>
<td>NAPT-Mathematics</td>
<td>3/Grade</td>
</tr>
<tr>
<td>Attendance</td>
<td>1/Grade</td>
<td>NAPK-Mathematics</td>
<td>3/Grade</td>
<td>Attendance</td>
<td>1/Grade</td>
</tr>
<tr>
<td>Promotion Rate</td>
<td>1/School</td>
<td>STEELS</td>
<td>1/Grade</td>
<td>SATI/ACT</td>
<td>4/School</td>
</tr>
</tbody>
</table>

* The weight columns show weight in the first part and whether the weight is applied to the variable at each grade or once for the entire school in the second part.
The final step involved multiplying the weight of each outcome variable by the standardized mean residual for each school, summing across variables, and obtaining a final weighted mean for each school. These means were then ranked to produce an ordering of effective schools.

Discussion

Dallas Independent School District (DISD) schools and their staffs were eligible for cash awards for the 1991-92 school year under the School Performance Improvement Awards program which was suggested by the Commission on Educational Excellence, a citizens group appointed by the Dallas Board of Education in 1990 to study the District's educational system. This recommendation was influenced by a similar program implemented in the DISD during the 1984-85 school year. By October 15, 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 for the program; the other half came from the community. Once a school had been selected as an award winner, it received $2,000, each member of its professional staff received $1,000, and each member of its support staff received $500. The equations discussed in this paper are those that were used for identifying outstanding schools.

One of the results of this program has been the creation of an Accountability Task Force. This twenty-four member Task Force consists of ten District teachers, four District principals, three District parents, five community representatives, and two central staff members. In addition, representatives of all of the various professional and support organizations, as well as the Parent-Teacher Association, are ex officio members of the Task Force. This Task Force is charged with the responsibility of overseeing the accountability system and specifically determining which outcome variables to use, which statistical model to use, how to assign weights to variables, and which schools, based on established criteria, are eligible for awards. The Task Force meets once a month and has been extremely involved in making informed decisions about the process. It has been an extremely positive outcome of the School Performance Improvement Awards program.

Besides providing an objective procedure for identifying effective schools, the program has a number of practical advantages. First, and most important, it is designed to foster teamwork among school staffs within schools. In order to achieve the necessary improvements in student outcomes school staffs must work together in a coordinated effort. The program does not reward individual competition among teachers within schools.
Second, the program focuses attention on the important outcomes of schooling. The Accountability Task Force, as well as many other groups associated with the schools, are discussing what it is that the schools are about. The process of weighing the outcome variables, a procedure that will be done annually, gives many divergent groups the opportunity to share their views relative to the purposes and importance of schooling. While the accountability system alone will not improve instruction, the curriculum and instructional delivery processes that must be changed to impact the defined outcomes will.

Third, the procedures described afford all schools an opportunity to be distinguished in the awards independently of where their student populations lie on the achievement continuum. The emphasis is on effectiveness with the students who come in the door, not absolute outcome levels. The techniques reward those schools that impact the most students the most positively.

Several alternate models are currently being explored for future applications of this program, although, at this point there is no reason to believe that the procedures described in this paper are not the most appropriate. Nevertheless, the authors would be less than thorough if they did not at least allude to these procedures.

Early in the process of choosing models, a time-series approach was considered. This model would have involved the use of time-series analysis on historical trends to establish predicted scores for each outcome variable. Once the extent of population reduction related to using more than one year of historical data was demonstrated, (about 29% per year) time-series analysis was deemed not to be a viable alternative.

Another approach that was considered is canonical correlation. This model requires the use of canonical correlation to establish a linear combination of dependent variables from which average deviation levels can be established for each school. In essence, the average deviation from the first canonical variate for the dependent variables is determined for each school. This procedure was found lacking due to the inability of the Accountability Task Force to individually weight outcome variables thus removing a major advantage of the program as discussed above. That is, canonical correlation, in reducing both the predictor and dependent variable space to essential combinations of variables may eliminate effects of variables which are perceived as critical by parents, community members, or practitioners.

Finally, hierarchical linear modeling was considered. This model requires the use of multilevel equations to define school effect. When attempting to apply this model in a
conventional manner, the investigators immediately encountered degrees of freedom problems in the within school equations. Without a means of adequately controlling this problem, it is unlikely that hierarchical linear modeling will provide a more effective model than the one currently proposed.
References


