DOCUMENT RESUME

ED 303 477 TM 012 528

AUTHOR Mandeville, Garrett K.; Heidari, Khosrow
TITLE Measuring School Effectiveness Using Hierarchical Linear Models.
PUB DATE Apr 88
NOTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Educational Assessment; Elementary Education; *Elementary Schools; Mathematics Achievement; *Measurement Techniques; Reading Achievement; Reliability; *Research Methodology; *School Effectiveness
IDENTIFIERS *Effective Schools Research; *Hierarchical Linear Modeling

ABSTRACT

Two major groups of researchers focus on identifying schools that have been unusually effective in terms of their students' achievement: (1) "effective schools" researchers; and (2) those charged with the responsibility of identifying schools for special recognition. However, all legitimate attemps to operationalize school effectiveness that are based on student achievement (i.e., those controlling relevant extraneous variables) have produced "measurements" that do not have desirable properties. For example, these measurements are quite unstable across different school subpopulations (such as students in different grades) and across years. This finding is counter-intuitive. In this study, eight alternative methods were used to "measure" the effectiveness in reading and mathematics of a sample of 135 elementary schools in South Carolina across grades 1 through 4 for the 1985-86 and 1986-87 school years. The eight student samples ranged from 8,000 to 11,000 children. Scores from the Basic Skills Assessment Programs and the Cognitive Skills Assessment Battery served as pretest measures. Four of the methods might be considered traditional, whereas the other four do not seem to have been considered previously. The results indicate that one of the new methods, based on hierarchical linear modeling, produced school effectiveness indices that were slightly more stable across grades and considerably more stable across years than the more traditional methods. Although several technical and political issues remain unresolved, this approach to the identification of exceptional schools should receive further consideration. Three tables of correlations are included. (Author/SLD)

Measuring School Effectiveness Using Heirarchical Linear Models

Garrett K. Mandeville

and

Khosrow Heidari

University of South Carolina

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 1988.

Abstract

There are two major groups of researchers who focus on identifying schools which have been unusually effective as regards the achievement of their students. One group are the "effective schools" researchers; the other are those charged with the responsibility of identifying schools for special recognition. Unfortunately, at this time all legitimate attempts to operationalize school effectiveness which are based on student achievement, i.e., those which control relevant extraneous variables, have produced "measurements" which do not have desireable reliability properties. For example, they are quite unstable across different school subpopulations (such as students in different grades) and across years. This "truth", which clearly is counter-intuitive, is slowly becoming recognized and its implications, especially for "effective schools" researchers, is very serious. In this research, eight alternative methods were used to "measure" the effectiveness in reading and mathematics of a sample of 135 schools across grades 1-4 for two consecutive years. Four of the methods might be considered traditional whereas the other four do not appear to have been considered previously. The results indicate that one of the new methods, based on heirarchical linear modeling (HLM), produced school effectiveness indices (SEIs) which were slightly more stable across grades and considerably more stable across years than the more traditional methods. Although there are a number of technical and political issues which are unresolved, this approach to the identification of exceptional schools should receive further consideration.

Measuring School Effectiveness Using Heirarchical Linear Models

For a number of years, researchers have attempted to examine how well individual schools have done in their efforts to foster important educational outcomes in the children who attend them. These studies have frequently utilized student achievement test scores and the focus has been issues such as school accountability, "school effectiveness", and the more recent efforts to award schools whose students have exhibited exceptional achievement. In fact, a recent National Invitational Conference on School Recognition Programs attests to the growing interest in this last issue.

Irrespective of the purpose for which school effectiveness indices (SEIs) are obtained, implicit in their computation is the notion that the resulting measures are stable. One of the problems in demonstrating this stability is that there are almost as many approaches to estimating SEIs as there are researchers studying the issue. (See Rowan, Bossart & Dwyer, 1983, for four general strategies and Frederick & Clauset, 1985 and Frechtling, 1982, for a variety of individual approaches.) Although a few researchers may still employ approaches based on absolute standards--a carry over from the definitions used in the early effective schools literature--most of the recent work has been based on regression models.

The studies which have been conducted to address the stability issue have also generally employed regression models; most have used the school as the unit of analysis with some appropriate standardization of the residuals serving as the SEIs. It should be noted that in many of the early studies the "school" in SEI is a misnomer since data for a single grade--and often a single subject--in each school were used. For example, Dyer, Linn and Patton (1969) studied 8th graders and Marco's (1974) research involved 3rd graders. These studies illustrated that, when based on subsamples of children in the same grade

during the same year, SEIs are reasonably stable (coefficients in the .7 to .9 range). It is important to realize, however, that these random subsamples of students received instruction from the same sets of teachers.

When Mandeville and Anderson (1987), addressed the issue of the stability across grades 1-4 (within the same school), they found small correlations in both reading (median r of .06) and mathematics (median r of .13). Matthews, Soder, Ramey and Sanders (1981), also reported that SEIs were quite inconsistent across grades 2-8. The implications of these findings for the identification of exceptional schools are clearly quite negative.

Nor do schools, or more properly the performance of students in a given grade in a school, demonstrate a high degree of consistency year-to-year. Forsyth (1973), studying 12th graders, found correlations which ranged from .11 to .50 (median = .28) depending on the subject area. Matthews et al. (1981) found that year-to-year correlations of SEIs computed at the same grade level ranged from -.24 to .44. Mandeville (1987) found correlations between .34 and .65 for reading and math SEIs in grades 1-4.

Interestingly, year-to-year consistency is an issue which may separate effective schools researchers and those interested in designing reward programs. It stands to reason that year-to-year consistency is--or at least should be--important for those desirous of studying schools identified as "effective" in an attempt to discern the correlates of effectiveness. Good and Brophy (1986) make this point in an analogy concerning studying the Super Bowl winner the following year, during which the team may not even make the playoffs. On the other hand Wynne, a leader in the school recognition movement, has recently recommended that schools be disallowed to repeat as winners (at least for a few years) to guarantee that a large number of schools receive awards (see Wynne, 1987). Of course Wynne didn't have the instability issue in mind when making this recommendation.

In most of the studies cited above, SEIs have been defined in a manner

consistent with Dyer et al. (1969), i. e., school level regression analyses have been applied to longitudinal data. The use of matched student data has been recommended as the most defensible by Dyer (1970) and Good and Brophy (1986), the rationalization being that this assures that the school can legitimately be held responsible for the students' performance. Thus the question must be raised as to whether 1) schools are really as unstable as the data suggest or 2) the results may be at least partially explained by the methodologies used. In this vein, Rowan (1985) has stated that the use of residuals from regression models is clearly an inappropriate way to operationalize SEIs. Rowan's point is well taken: it is probably time to investigate other ways to measure school effectiveness.

One promising new analytic technique is the use of heirarchical linear models (HLM). For example, Raudenbush and Bryk (1986) indicate that the Empirical Bayes (EB) estimates of the within school effects produced by their operationalization of HLM have smaller mean squared error than OLS estimates and the suggest using them for the purpose of identifying "unusually effective schools." (p. 14). This paper will investigate the grade-to-grade and year-to-year consistency of SEIs operationalized according to four more-or-less traditional approaches and four which do not appear to have been used previously. Two of the untried methods will employ HLM.

Eight Methods of Defining SEIs

Given the availability of data, most reasonable operational definitions of SEIs control for the ability level or achievement of the student prior to the period for which the assessment is to apply. Let us refer to this variable as the pretest (X). Another important variable to consider is the socio-economic-status (SES) of the student population. Finally, each child must be tested subsequent to the period of instruction which is being assessed and this will be referred to as the posttest (Y). In this research, it is assumed that the three variables-X, Y and SES--are available for each student who was in

continuous residence in the schools under consideration. Furthermore, it is assumed that X and Y may, as is likely, be in different metrics for students at different grade levels and for different subject areas tested. Although other variables--if available--should also be considered, this will be the setup in the research reported in this paper.

Obviously, there are a number of ways to combine this information to assess a school's "performance." One popular approach is to aggregate all three variables to the school level, i.e., compute the school means, and then regress Y onto X and SES. Some function of the residuals from the regression surface is then used as a measure of the performance of the students at that grade level in the subject area (e.g., reading) of the test.

There are, however, at least two ways to conduct the regression analysis not to mention the possible functions of the residuals. By this we are referring to whether an unweighted or weighted least squares (WLS) approach is used in the regression analysis. An unweighted analysis appears to be more common and is clearly the better of the two from a political standpoint. A likely rationale for an unweighted analysis is that each school should be represented as 1/Kth of the population of K schools, irrespective of the number of children attending the school. From this viewpoint standard errors associated with the school means are irrelevant since no sampling has taken place, i.e., the the school means are considered to be parameters.

Of course proponents of a WLS analysis must take the opposite side of the above arguments, a position which may be difficult to defend. It also may be difficult to explain to educators and the lay public that the regression function is "pulled" toward the data points based on the larger schools. Also, assuming that the SEIs are based on standardized rather than raw residuals, the standard errors associated with the residuals for the smaller schools are smaller than those associated with larger ones. This means that the actual performance of a low enrollemnt school must be further above (or below) its

predicted performance level than is true for a larger school in order to achieve a comparable grade-subtest specific SEI. Although these arguments appear to be difficult to counter, a WLS solution will be explored to see if the resulting SEIs are more stable.

In the two approaches described above, school aggregates were formed prior to the regression analysis. An alternative way to use the available information is to conduct the regression analysis at the student level and then aggregate the standardized residuals to the school level. (See O'Connor, 1972 for some technical arguments against this approach.) Once again, if one takes the position that the actual children available for a given school represent a sample from some hypothetical population, the question of weighting can be raised. A weighted analysis would require that the mean residual be standardized by multiplying it by the square root of n. Again, this approach will be considered in this paper.

In the four approaches described above, the available data were utilized either at the school level or the student level. This was possible because it was assumed that X, SES and Y were measured at the student level. A purely student level analysis would have been impossible, however, if we had desired to control for school level variables which were not aggregates of student variables (e.g., average teacher salary). Only the first two models discussed above, which employed school means, would accomodate the inclusion of such contextual variables.

There is a middle road, however, namely to first model within each unit (here school) and then use the obtained estimates as the dependent variables in a between units model. This has been the approach in the so-called "slopes-as-outcomes" research. In the context of obtaining SEIs, however, the focus is on intercepts rather than slopes and this suggests that the within school predictor(s) should be centered. If centered at the school mean(s), then the intercept in the OLS solution, i.e., b_0, will be the school mean on Y; if

centered at the grand mean(s), then b_o is analogous to the adjusted mean in analysis of covariance. The latter result, which provides an estimate of average school performance if all schools began with students who were comparable on X, seems more appropriate for this application. (It does, however, assume reasonably overlapping distributions of the predictors across the schools being studied if the results are to be meaningful.) In the between schools model, b_o is the dependent variable to be predicted by the school level data.

In a logical extension of the traditional approaches, the residuals from this second model would serve as SEIs since they would reflect the amount of discrepancy between school performance (adjusted for within school predictors) and the prediction of that performance based on between school predictors.

The most straightforward way to fit these models is to use OLS at each stage and this was done to produce one solution. The HLM approach as described by Raudenbush and Bryk (1986), however, is to use an EB solution. Many of the advantages which they claim for the HLM approach (see, e.g., Raudenbush & Bryk, 1986, pp. 2-3) lose some relevance in this application since our interest is focussed on the intercept which is much less prone to sampling error than the regression coefficients which are usually of interest. The EB estimates of the intercept, however, do differ from the OLS estimates, although the differences are often quite small. Generally speaking, extreme values of b_o will be pulled toward more central positions based on the data and the amount of this "shrinkage" will be inversely related to the sample size. Thus, our fifth and sixth candidates will be the residuals from the HLM and OLS estimates of b_o in the between schools model.

Because residuals have proven to be so unstable, we will also employ two other statistics as potential SEIs. These will be the two estimates of b_o themselves. When OLS estimation is used, these will simply be the b_os from the between schools model, i.e., they will not have been adjusted for SES. For HLM

estimation, however, b_o is jointly adjusted for both X and SES. The EB and OLS estimates of b_o will serve as our seventh and eighth SEI operationalizations.

It should be noted that an assumption in the heirarchical analysis is that the regression coefficients in the within-units model are considered to be random variables for, otherwise, they could not serve as the dependent variables in the between-units model. This raises the question of the logic of using them to order or reward schools. The question becomes moot, however, when it is reconsidered that all of the traditional methods cited above are based on residuals, i.e., random errors, from regression models.

Instruments

For a number of years, South Carolina has tested all students in the majority of the grades in the K-12 span. Criterion referenced tests (CRT) used as a part of the Basic Skills Assessment Program (BSAP) are administered each spring to all students in grades 1,2,3,6,8, and 10. Students in grades 4, 5, 7, 9, and 11 are tested, also in the spring, with the Comprehensive Tests of Basic Skills (CTBS). In addition, the Cognitive Skills Assessment Battery (CSAB) is administered at the beginning of the 1st grade as a school readiness test. Except at grade 10, the BSAP reading and mathematics tests are relatively short (36 and 30 multiple choice items, respectively) and provide, among other metrics, scale scores.

Methods

This study was limited to the elementary schools in South Carolina which contained students in grades one through four during the 1985-86 and 1986-87 school years. (It is likely that most schools also had students at additional grade levels but, where present, these data were not considered.) Student records for the Spring 1986 and 1987 testings were matched with the corresponding test records for the most recent prior testing. Schools with fewer than 10 matched, complete student records at each of the four grade levels for both years were eliminated from consideration and, in order to

reduce the amount of computer time required to conduct the study, a systematic one-in-three sample of the schools satisfying these criteria was selected. The resulting school sample size was 135 and the eight student samples ranged from about 8,000 to 11,000 children. Both disaggregated and aggregated data sets were created for each grade-year combination.

BSAP scale scores in reading and math (grades 1-3), and expanded scale scores for the Total Reading and Total Math subtests of the CTBS (4th grade) based on the Spring 1986 and 1987 testings were used as the posttests (Y) for each of the four grade cohorts. Prior year BSAP scale scores (grades 2-4) or fall CSAB raw score (1st grade) and lunch status (dichotomized with students eligible for free or reduced price lunches in one category and those not eligible in another) served as X and SES, respectively. Previous reading scores were used to predict reading and previous math scores to predict math and the analyses were conducted separately for reading and math posttest data.

The four traditional SEIs were obtained by using the REG procedure of the Statistical Analysis System (SAS, 1985) with "studentized" residuals serving as the SEIs (school level models) or as the aggregation unit (student level models). The other four SEIs were obtained using Version 1.0 of the HLM computer program described in Bryk, Raudenbush, Setzer, and Congdon (1986).

For a given subject area, eight SEIs were produced for each grade level for each year. Results will be presented to indicate:

1. The degree of similarity of the eight approaches
2. The stability of each approach <u>across grades</u> for each subject area and year
3. The stability of each approach <u>across years</u> for each subject area and grade
4. The stability of each approach <u>across years</u> based on composites across subject areas, across grades, and across both subject areas and grades

Results

Preliminary Regression Analyses

Summary results of the regression analyses relevant to the traditional SEIs are reported in Table 1. As we have generally found with similar data,

---

Table 1 About Here

---

student (and group) performance in reading is generally more predictable than performance in math although the R^2s for math would have been somewhat larger if the reading pretest had been used as an additional predictor for grades 2-4. (At grade 1 the CSAB score is the only available "pretest".) In reading for grades 2-4 school means are more predictable than individual scores, which is also a common finding. Surprisingly, this effect does not occur in grade 1 for either subject area and is quite small in math at grades 2-4. The SES measure was not a significant predictor (in addition to pretest) in a few of the school level analyses and, when this occurred, it was generally not a significant predictor in the 2nd stage of HLM modeling.

Associations Among the Eight SEIs

The intercorrelations among the eight SEIs within each grade, subtest, and year were obtained. Since there is such a large number of rs--28 nonredundant values for each grade, subtest, and year for a total of 448--no tables will be provided. Rather, a few summary statements are in order.

The first general impression is that the rs are exceptionally large; the median across the whole set is about .95. Curiously, the rs for the SEIs at the 1st and 3rd grades were large almost irrespective of the basis for computation of the SEI with approximately 73% being larger than .95 and fewer than 3% less than .90. The correlations for grades two and four were not as large but 84% were .80 or larger. The smallest correlation was .56.

Among the traditional formulas, weighting (either the solution based on

means or aggregated student scores) changed the results only slightly as measured by these correlations. Generally the rs were in the .96 to .97 range with a few as low as .94 and as large as .99. There was a slight tendency for the rs between the unweighted and weighted approaches based on student level regression analyses to be larger than those based on school means. Once again, the exceptionally large values (beyond .98) tended to occur at the 1st and 3rd grades.

Among the less traditional SEI formulations, it is quite understandable that the associations were larger for the SEIs of the same "type." By this we mean that the 5th and 6th SEIs are essentially based on residuals, and the 7th and 8th SEIs are based on estimates of intercepts. All but three of these 32 rs (two sets of SEIs--the 5th vs. 6th and 7th vs. 8th--by 4 grades by two years by two subject areas) were .96 or larger and roughly two-thirds were at least .98. At grades one and three, however, the relationships among SEIs based on residuals and those based on b_o were even quite large, always beyond .90, whether the estimation was OLS or EB. For the other two grades, these associations were more modest, however, generally being in the .60 to .80s.

Finally, the nontraditional SEIs which were based on residuals were more likely to be related to the four traditional SEIs (all of which were also based on residuals) than those characterized as b_os. In fact, all the SEIs based on EB residuals correlated beyond .90 with the 1st four SEIs and most based on OLS estimation did likewise.

To summarize, the eight versions of SEIs appear to be in large part capturing the same aspects of the data and, for some inexplicable reason, the associations are consistently larger in grades one and three than in the other two grades; unweighted and weighted versions of the same basic SEI are highly correlated; nontraditional SEIs based on residuals, especially under EB estimation, are highly associated with the four traditional indices and, quite logically, the two nontraditional SEIs based on intercepts are less related.

These data would suggest that it is unlikely that any of the eight methods would prove advantageous from the standpoint of stability.

Cross-Grade Correlations

Next we will consider the intercorrelations among the grade-specific SEIs since it is their instability that is the most perplexing. For each year and subject area, the six intercorrelations among the four grade-specific SEIs (i.e., those for grade 1 vs. grade 2, grade 1 vs. grade 3, etc.) were computed. Summary statistics in terms of the median, the minimum and the maximum (across the six rs) for each combination of year and subject area and the same data for all 24 rs (the Total column) are reported in Table 2.

Table 2 About Here

The four traditional SEIs are seen to be unstable across grade, as expected, and these data could hardly be used to select the "best" of the four from this standpoint. The two nontraditional SEIs based on residuals are no improvement either; if anything, they appear slightly less stable on the average than the four traditional indices.

The results regarding OLS_B0 are slightly better; none of the rs is negative and they range from 0 to an almost respectable .35. Among the eight candidate SEIs, however, the results for EB_B0, although still quite unstable, appear to stand out. The rs range from .01 to .55 and, across the full set of 24 rs, 16 (67%) differ statistically from zero at the .05 level (r greater than about .19).

The results for EB_B0 (and also for OLS_B0, the other SEI which is based on an estimate of b_o) differ from the other six which are based on residuals, in one another respect (Hereafter, the first six will be collectively referred to as the RES methods and the last two as the INT methods.) Although endency is not great, Table 2 suggests that the RES SEIs are more stable

in math than in reading. Similar results have been found by Mandeville and Anderson (1987) among others. This does not appear to be true for the INT SEIs where the results in reading and math, although inconsistent, are reasonably comparable.

Cross-Year Correlations

The correlations between the SEIs based on the 1986 and 1987 data are presented in Table 3. In addition to the results for each grade and subject

---

Table 3 About Here

---

area, correlations based on various aggregation strategies are also included. Thus, correlations representing a simple averaging of the reading and math SEIs are referred to as RMAVE and those corresponding to an average across the four grade-specific SEIs are denoted as COMP for composite. These were included since they are often computed to obtain more comprehensive measurements of the performance of the students in a school. For example, COMP in the 3rd set of columns (RMAVE) represents the results under a double-averaging scheme which incorporates eight data points.

With a few minor exceptions, the year-to-year rs across the RES SEIs are quite consistent. For this group, 1st grade reading is most stable (rs in the .60s) and the correlations for reading at the other grade levels and math at all grade levels range from about .30 to .50. Although we might bend to a accept a stability coefficient in the .60s as a minimum standard of consistency, clearly these latter results cannot be characterized as describing schools that are being measured in any consistent fashion. Nor do various compositing strategies produce more stable year-to-year measurements for these RES SEIs since the largest r for the most comprehensive one, the double-average for UNMEAN, was only .45.

Once again, the results for the last two SEIs, do not "track" with those

of the first six. For reading, the "intercept" based SEIs are quite stable for all but the 3rd grade and the r for the reading composite is respectable at .67 for EB_B0. In math, on the other hand, these SEIs are only slightly more stable than the RES set. Thus, although the r of .51 for the EB_B0 math composite is the largest such value, it is only trivially larger than the .49 for UNMEAN. Under double-compositing, however, EB_B0, with an r of .62, is clearly the most stable. The OLS_B0 SEI tracks EB_B0 but is slightly less stable year-to-year.

Summary and Discussion

The objective of this study was to investigate the possibility that new operationalizations might produce SEIs which could legitimately be called measurements. As the title indicates, the emphasis was to be on the relatively new analytic procedure HLM; however, modifications of some of the more traditional approaches were also considered.

The four more-or-less traditional and two of the new methods are based on residuals from regression analyses. These six SEIs were highly intercorrelated. The consistency of the stability (or lack of stability) of these six SEIs was surprising; the results are very resistant to change. It seems to make little difference which aggregation unit is chosen, whether a weighted or unweighted solution is used, or which estimation algorithm is employed. Cross-grade rs tend to be less than .1 in reading, and slightly larger in math (the median of the medians was about .14). With the exception of 1st grade reading, cross-year correlations were usually in the .30s of .40s and various compositing strategies did not increase stability. Thus we have more evidence in support of Rowan's (1985) contention.

The two SEIs based on estimating the intercept in a regression model performed slightly better as regards inter-grade stability and considerably better as regards year-to-year stability. (This was somewhat surprising since correlations among the six RES SEIs and the two INT SEIs were quite large.)

Of the two INT approaches, the one based on EB estimation was clearly superior. In fact, some might be satisfied with the year-to-year stability of .62 for the average over subject area and grades using the EB_B0 approach.

The focus on stability across grades and over years was a very limited one and this research is clearly not definitive for a number of reasons. For one thing, inventing an SEI which is relatively stable over these dimensions proves nothing since it is almost certainly true that a number of stable but inappropriate SEIs could be constructed. For example, means on achievement test scores--with no attempt to control for the factors which affect them--are likely quite stable. Another drawback associated with the EB_B0 SEI is that it would be difficult to communicate how it is computed to the affected educators and lay public which is probably necessary for a school reward program. Finally, it is likely that many of the objections noted above concerning weighted solutions also apply to the EB estimation approach. If so, it might not be politically feasible to suggest it as an approach to be used for rewarding schools but might suffice for effective schools researchers.

On the other hand, the EB_B0 SEI definitely has sufficient promise to warrant further study. One of the issues which must be addressed is the mechanism by which this approaches achieves greater stability. Also, different models which include more comprehensive sets of predictors also need to be considered to ascertain whether stability coefficients can be increased to more acceptable levels. In this study a very limited set of variables (only SES) was included in the second stage model; clearly a number of potentially relevant additional variables should be considered. Also, the OLS_B0 SEIs exhibited stability which, although somewhat less than the EB_B0 SEIs, was comparable. Since OLS_B0 may be more easily obtained using commonly available software packages, this approach should also receive further investigation.

Furthermore, time is of the essence. It is particularly important that, better methods of defining SEIs exist, they be quickly identified and

communicated to the interested parties, i.e., the "effective schools" researchers and those charged with the responsibility of identifying schools for special recognition. This, in spite of the fact that it is not clear that researchers in either group realize the seriousness of the instability issue.

This research has brought a modicum of evidence to bear on a "truth" which has seemed self-evident to educators for quite some time, namely, that some schools are better than others in a fairly global sense. These results suggest that this belief may contain more validity than has heretofore been demonstrated. Although year-to-year stability coefficients in the .60s and especially cross-grade coefficients in the mid .20s are not as high as we would like, they are a step in the right direction.

References

Bryk, A. S., Raudenbush, S. W., Seltzer, M. & Congdon, R. (1986) An Introduction to HLM: Computer Progran and User's Guide.

Dyer, H. S. (1970). Toward objective criteria of professional accountability in the schools of New York City. Phi Delta Kappan, 52, 206-211.

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

Forsyth, R. A. (1973). Some empirical results related to the stability of performance indicators in Dyer's student change model of an educational system. Journal of Educational Measurement, 10, 7-12.

Frechtling, J. A. (1982). Alternative methods for determining effectiveness: Convergence and divergence. Paper presented at the annual meeting of the American Educational Research Association, New York.

Frederick, J. M. & Clauset, K. H. (1985). A comparison of the major algorithms for measuring school effectiveness. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Good, T. L. & Brophy, J. E. (1986). School Effects. In M. C. Whittrock (Ed.), Handbook of Research on Teaching (3rd ed., pp. 570-604). New York: Macmillan.

Mandeville, G. K. (1987). The stability of school effectiveness indices across years. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington DC.

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

Marco, G. (1974). A comparison of selected school effectiveness measures based on longitudinal data. Journal of Educational Measurement, 11, 225-234.

Matthews, T. A., Soder, J. B., Ramey, M. C. & Sanders, G. H. (1981). Use of district test scores to compare the academic effectiveness of schools. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

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

Raudenbush, S. & Bryk, A. S. (1986). A heirarchical model for studying school effects. Sociology of Education, 59, 1-17.

Rowan, B. (1985). The assessment of school effectiveness. In Kyle, R. M. J. (Ed.). Reaching for excellence: An effective schools sourcebook. N. I. E.: Washington DC.

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

SAS Institute, Inc. (1985). SAS User's Guide: Statistics, Version 5 Edition. Cary, NC: Author.

Wynne, E. A. (1987). Characteristics of an ideal recognition program. An address delivered at the National Invitational Conference on School Recognition Programs, Miami.

Table 1
Squared Multiple Correlations For Various Regression Models
By Year, Subject Area and Grade Level

Grade	Model	Reading 1986	Reading 1987	Mathematics 1986	Mathematics 1987
1	UNMEAN	36	44*	26	27*
	WTMEAN	39	46	31	32*
	UNSTUD	39	40	32	29
2	UNMEAN	73	68	39	38
	WTMEAN	76	71	47	40
	UNSTUD	47	41	22	20
3	UNMEAN	60	62*	29	26
	WTMEAN	65	68	35	29
	UNSTUD	44	43	24	25
4	UNMEAN	83	78	45	44
	WTMEAN	85	80	47	47
	UNSTUD	58	54	37	34

Note: UNMEAN is the unweighted analysis of school means, WTMEAN is the weighted least squares analysis of school means, and UNSTUD is the unweighted individual (student) analysis; asterisks (*) indicate that the SES variable was NS at the .05 level of significance. Leading decimals have been omitted.

Table 2
Median, Minimum and Maximum Cross-Grade Pearson Correlations
By Year and Subject Area
For Eight SEIs

SEI	Reading 1986	Reading 1987	Mathematics 1986	Mathematics 1987	Total
UNMEAN	10(-01,24)	08(01,22)	23(11,26)	14(05,20)	13(-01,26)
WTMEAN	06(-07,20)	09(01,19)	18(03,29)	14(08,18)	13(-07,29)
UNSTUD	08(-11,21)	06(-05,23)	17(00,26)	08(00,16)	09(-11,26)
WTSTUD	07(-14,16)	08(-04,20)	14(-06,33)	10(01,14)	10(-14,33)
EBRES	00(-17,19)	04(-05,21)	19(-06,23)	05(-02,14)	04(-17,23)
OLSRES	02(-16,17)	05(-02,24)	13(01,26)	07(-04,14)	05(-16,26)
EB_B0	24(13,55)	24(14,52)	21(12,37)	16(01,26)	22(01,55)
OLS_B0	17(08,32)	15(09,35)	23(12,27)	15(00,22)	15(00,35)

Note: leading decimals have been omitted. The first three mnemonics are the same as those defined in Table 1 above; the others are:
WTSTUD - the weighted analysis using student residuals
EBRES - residuals from the 2nd stage HLM estimates of B_o
OLSRES - residuals from the 2nd stage OLS estimates of B_o
EB_B0 - 2nd stage HLM estimates of B_o
OLS_B0 - 2nd stage OLS estimates of B_o

Table 3
Cross-Year Pearson Correlations
By Grade and Subject Area
For Eight SEIs

	Reading					Mathematics					RMAVE				
SEI	1	2	3	4	COMP	1	2	3	4	COMP	1	2	3	4	COMP
UNMEAN	63	30	28	37	34	44	39	46	53	49	57	35	42	45	45
WTMEAN	64	31	28	40	36	43	43	41	49	44	57	38	40	45	43
UNSTUD	63	36	30	43	31	45	40	43	52	39	58	39	43	49	39
WTSTUD	64	38	31	48	36	44	45	39	47	35	59	43	42	49	39
EBRES	63	29	27	41	33	44	38	41	46	35	58	35	41	44	37
OLSRES	61	28	25	34	29	47	38	42	44	36	58	35	42	38	36
EB_B0	65	63	34	71	67	44	54	45	59	51	59	61	46	69	62
OLS_B0	62	50	28	55	55	47	48	46	55	48	59	52	45	57	55

Note: leading decimals have been omitted.