# Instructional Variation and Student Achievement in a Standards-Based Education District

**CSE Technical Report 522**

Lauren B. Resnick and Michael Harwell
CRESST / University of Pittsburgh

June 2000

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1 Construct Validity: Understanding Cognitive Processes—Performance Tasks in Standardized Tests Lauren Resnick, Project Director CRESST/University of Pittsburgh, Learning Reseach and Development Center



The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

# INSTRUCTIONAL VARIATION AND STUDENT ACHIEVEMENT IN A STANDARDS-BASED EDUCATION DISTRICT

**Lauren B. Resnick and Michael Harwell**

**CRESST/University of Pittsburgh**

## Abstract

This paper, part of a larger study of the links between instructional variation and variation in performance on standards-based assessment, reports on the relations between examination results and instructional variation in a diverse New York City school district. The district has put in place an educational improvement system founded on intensive, school-based professional development that is carefully related to a preferred framework for teaching literacy. This study provides strong evidence that the school district's school-based professional development program improves teaching quality for diverse schools in ways that affect students' achievement scores.

This paper reports the second in a series of investigations of the relations between instructional variation and variation in performance on standards-based assessments. A standards-based strategy of school reform is a means to establish a more equitable system of education by shifting from an aptitude-based to an effort-based mode of operation (e.g., Howard, 1995; Resnick, 1995). An effort-based educational system has five key elements:

- Clear expectations for achievement, well understood by everyone;
- Fair and credible evaluations of achievement;
- Celebration and payoff for success;
- As much time as necessary to meet learning expectations;
- Expert instruction.

The clear expectations criterion of an effort-based system calls for teachers and students to know clearly what kinds of learning are expected and to direct their teaching and learning energies in a targeted way to meeting standards that will matter in their lives. Closely linked to the principle of clear expectations is the use,

within the system, of fair and credible evaluations: tests and other assessments that are well aligned to the clear expectations articulated in the system's standards. Fairness demands that students can study for these assessments and teachers can legitimately prepare their students for them. Students and teachers need recognition of their work as they prepare for these assessments. The use of linked standards and assessments, often with attached consequences, creates a virtual right both to expert instruction for students and to as much instructional time as necessary to meet the standards. Because teachers in such a system are likely to be called on to teach in new ways, teachers also need continuing and expert professional development.

In an earlier paper in this series, Yoon and Resnick (1998) showed that teachers who participated in the professional development activities of the California Mathematics Renaissance taught differently from a comparison group of teachers and that their students had significantly higher levels of performance on key portions of the New Standards Reference Examinations (Harcourt Educational Measurement, 1997, 1998, 1999), which systematically measured the kinds of knowledge and skill stressed in the Renaissance program. Furthermore, there was a smaller performance gap between white and minority students in the Renaissance group than in the comparison group, controlling for socioeconomic status. These results constitute a weak but positive test of the effort-based education strategy. They show that professional development stressing content and teaching methods aligned with a high-demand assessment somewhat improved student performance and especially benefited minority students. But the study used only an indirect measure of instructional quality (teacher and student self-reports). Furthermore, because of the structure of the California Mathematics Renaissance—a statewide program of professional development that enrolls individual teachers on a voluntary basis and is not linked to any specific district's teaching program—the study could not examine the effectiveness of a more tightly structured effort-based system.

This paper explores relations between examination results and instructional variation in a school district that has put in place an educational improvement system founded on intensive, school-based professional development. The district in question is Community School District 2 in New York City. That district, in Manhattan, has an exceptionally varied population of elementary and secondary school students. Its 26 elementary schools, the focus in the present study, include 6 schools in which at least 90% of the students are eligible for free or reduced-price

lunches and 6 others in which less than 21% are eligible. Some of the latter schools are located in neighborhoods that have among the highest per capita incomes in the United States, and the district has successfully attracted many middle-class children in an environment peppered with private schools. Despite these variations among schools, the district's leadership espouses a vigorous theory of educational equity and aims for high-quality instruction in even the most socioeconomically challenged schools.

The District 2 theory and practice of educational management via professional development is described in papers by Elmore and Burney (1997a, 1997b), Fink and Resnick (2000) and Resnick and Hall (1998). That theory calls on school principals to organize and lead learning opportunities for their teachers, often calling on a district-organized network of instructional coaches. Professional development is usually carefully related to a preferred framework for teaching literacy (see Maloy, 1998a, 1998b; Stein & D'Amico, 1998, 1999). Teachers recognized as expert in particular strategies of teaching within this literacy framework serve as coaches and mentors to others within their own schools and in other schools.

In a series of interviews, the then-Deputy Superintendent[1] and the Director of Professional Development[2] of District 2, further articulated their theory-of-action for increasing student achievement. The District 2 theory of action can be schematized as a hypothesized model of school improvement, which is shown in Figure 1. The arrows indicate the expected direction of the effect of one variable on another. Focusing first on variables that educators can directly influence, student achievement is hypothesized to be a function of teaching quality, which is hypothesized to be a function of the quality of staff (teachers) and the extent and quality of professional development activities in the school. Quality of staff and professional development, in turn, are a function of the school principal's leadership.

Teaching quality, quality of staff, and professional development are the variables that are the focus of the district's policies and practices, the variables that district leadership believes educators can act on directly. They also recognize,

[1] Elaine Fink, who was Deputy Superintendent at the time this study was conducted, is now Superintendent of District 2.

[2] Beatrice Johnstone, who was Director of Professional Development at the time this study was conducted, is now one of two Deputy Superintendents of District 2.

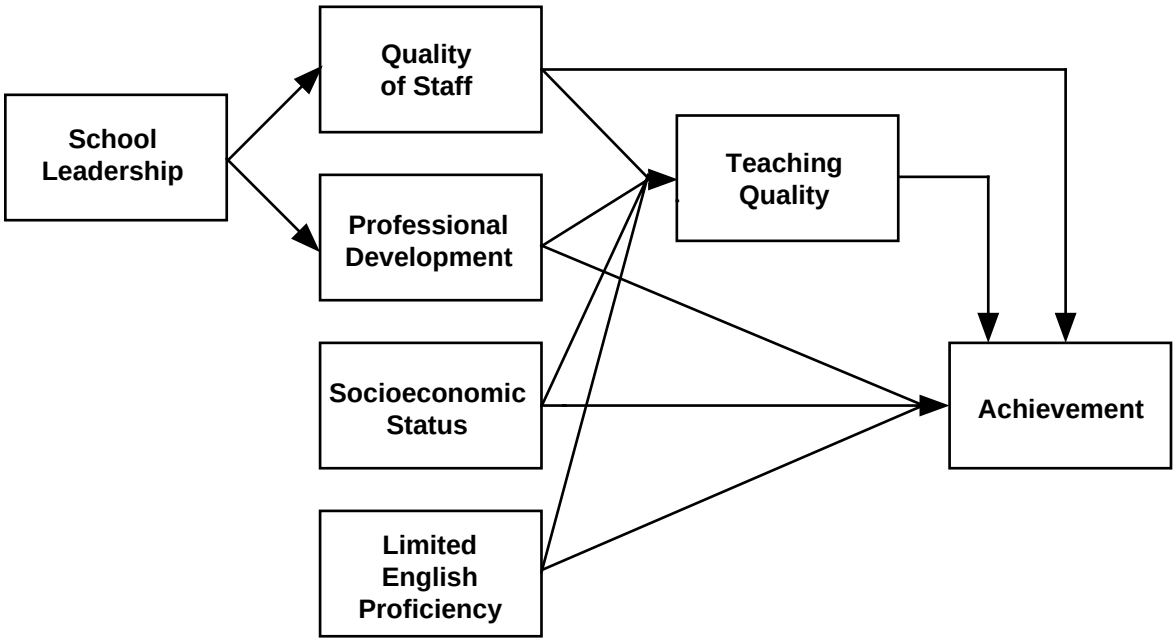


*Figure 1.* District 2 theory of action for increasing student achievement.

however, that the population of students in a school will affect achievement scores, both directly (because children learn at home as well as at school) and indirectly (because it is harder to recruit and retain teachers in schools with high proportions of limited English proficient (LEP) and low socioeconomic status (SES) children).

Elements of the District 2 theory are shared by many education practitioners and scholars. Only a few districts, however, have made school-based leadership and professional development the centerpiece of a strategy for improving educational equity and opportunity to learn. Our study examines the extent to which District 2's actual practice and its achievement results confirm the theory.

## Method

The District 2 theory of action, shown in Figure 1, was used to guide our investigation. Because the District 2 theory of school improvement focuses on the principal's responsibility for teaching quality in the school, we used schools in the district as the unit of analysis. For each school, SES was measured as the proportion

of students who were *not* eligible for free or reduced-cost lunches, and language status was measured as the proportion of students classified officially as limited English proficient (LEP). Measures of School Leadership, Professional Development, Quality of Staff, and Teaching Quality were derived from ratings made by the district's deputy superintendent and director of professional development. Ratings were also made on several other features of schools, including relations of the school with parents and community, that are not central to District 2's theory of school improvement but are held by other theorists to be important. Student achievement was assessed using the New Standards Reference Examinations in Mathematics and English Language Arts, which were administered in all District 2 schools in spring 1997.

**Student Achievement Measures**

The New Standards Reference Examinations are performance and multiple-choice on-demand assessments that are systematically referenced to the New Standards Performance Standards for Mathematics and English Language Arts (New Standards, 1997). New York City has adopted these standards, and District 2 is implementing them as a key part of its high-performance learning strategy. Scores on the New Standards exams, therefore, should reflect the status of instruction and professional development efforts in the district. Information from students who did not complete the New Standards exams, which are typically administered over three days, was not used in any data analyses.

New Standards scores are reported in standards clusters; each student receives a grade on each cluster. Students received grades in four English Language Arts standards clusters: Reading–Basic Understanding; Reading–Analysis and Interpretation; Writing; and Conventions. They also received grades in three Mathematics clusters: Mathematical Skills; Conceptual Understanding; and Problem Solving. In each cluster, a student receives one of five grades: Achieved the Standard with Honors; Achieved the Standard; Nearly Achieved the Standard; Below the Standard; and Little Evidence of Achievement.

The data available for each school consisted of the percentages of students receiving each grade in each standards cluster. For ease of reporting and interpretation, we combined the two top grades (Achieved the Standard and Achieved the Standard with Honors) for each cluster into a single score for each school, called *Percentage of Students At or Above Standard*, and the two bottom grades

(Below the Standard and Little Evidence of Achievement) into a second score, called *Percentage of Students Well Below the Standard.*

For purposes of regression and path analyses, the achievement scores were further aggregated, based on the high correlations between the cluster scores within English Language Arts and Mathematics. A principal components analysis for percentage of students scoring at or above standard for the four English Language Arts standards clusters produced one clear component that accounted for 92% of the variance. A similar analysis for percentage of students well below the standard on the four English Language Arts standards clusters also yielded one component, which accounted for 91% of the variance. Parallel analyses for Mathematics yielded one component for percentage of students above the standard (accounting for 93% of the variance) and another for percentage of students well below the standard (accounting for 94% of the variance). Accordingly four aggregate variables—EnglishAbove, EnglishBelow, MathAbove, and MathBelow—were computed by averaging the cluster scores.[3]

**School Quality Ratings**

School-level measures of professional development, quality of staff, and quality of teaching were generated from ratings of the district's 26 elementary schools made by the Deputy Superintendent (Elaine Fink) and the Director of Professional Development for the district (Bea Johnstone). The use of ratings by insiders to the system, instead of more objective judgments by researchers or practitioners without a vested interest in the theory being tested, is unusual for research of this kind and calls for some explanation. We could have sent external visitors to the schools and classrooms under study, but their judgments would have been based on far less information about the schools and classrooms than those that Fink and Johnstone could provide. These senior administrators in District 2 make frequent visits to each of the schools and are intimately familiar with the quality of teaching and the level of student work in classrooms. They possess a deep knowledge of the instruction in each school. We chose the latter course, and our results (see footnote 4)—confirming some, but not all, details of District 2's theory of action—provide evidence that there was adequate objectivity in our procedure.

[3] For example, the EnglishAbove score for a school was computed by summing the percentage of students at or above standard on Basic Reading, Reading Analysis, Writing and Conventions and dividing by four.

No established measures for most of the variables in this study exist, and the process of developing rating scales began with extended interviews of Fink and Johnstone, conducted by Lauren Resnick. On the basis of these interviews, a set of 13 features on which schools could be rated were established. These features are described in Table 1. The rating features reflect the District 2 theory represented in Figure 1, but for most of the components of Figure 1 more than a single rating was made because these sophisticated practitioners made distinctions in judging practice that were finer-grained than those expressed in their overall theory of action. Thus, two separate ratings of teaching quality were included, one based on judgments of the quality of student work (variable #1) and one based on observations of the teachers' interactions with students (#2). Other rating variables represented judgments of quality of staff (#3) and professional development in the school (#4).

Table 1

Rating Features

| |
|---|
| #1-Quality of Student Work: Judgements based on observations at the school, not on test scores. |
| #2-Quality of Teaching: Judgements of the overall quality of classroom instruction observed in the school. |
| #3-Quality of Staff: Judgements of the overall quality of teaching staff in the school. |
| #4-Professional Development: Judgments of the extent and quality of professional development activities in the school. |
| #5-Parents and Community: Judgments of the degree to which the school is connected to parents and to the community. |
| #6-Leadership–Culture: Extent to which the principal has established a culture of continuous professional development and improvement among staff. |
| #7-Leadership–Content: Extent to which the principal focuses professional development on instructional practice and content. |
| #8-Leadership–Discriminate: Extent to which the principal can discriminate quality of teaching. |
| #9-Leadership–Select: Extent to which the principal has selected new teachers well. |
| #10-Leadership–Weed: Extent to which the principal has been successful in weeding out very weak teachers. |
| #11-Improvement: Degree of improvement in the school since the current principal came on board. |
| #12-Global: Judgments of the overall quality of the school. |
| #13-Potential: Judgments of the extent to which the school is now poised for improvement. |

Five separate ratings of principals' leadership capacities were made. Two focused on the principal's leadership of professional development in the school, with one rating the principal's ability to establish a positive culture of professional sharing in the school (#6), and the other rating the principal's ability to actually lead teachers in understanding the content of instruction and pedagogy (#7). Three leadership variables had to do with the principal's capacity to assemble and retain a high-quality staff. These ratings (#s 8, 9, and 10) reflect the principal's ability to distinguish strong from weak teaching, to select good teachers when openings developed, and to find ways to weed out extremely weak teachers.

Three additional variables (#s 11, 12, and 13) were global judgments by Fink and Johnstone of the quality of each school. Variable #5 (parents and community) was included in the rating list at the suggestion of the investigators, although it is not an articulated part of the District 2 theory of action. We asked for its inclusion because so many other theorists view it as an important contributor to successful efforts to raise student achievement, and we wanted to include some test of its role in District 2.

Once the features were defined, Fink and Johnstone independently rated each elementary school in the district on each variable, using a scale of 1 to 10, with a 10 rating being the highest positive judgment. Their ratings were made without knowledge of the New Standards examination results, although they were familiar with the history of each school's performance on state- and city-mandated norm-referenced tests. Disagreements were discussed, and in most cases, a common rating was agreed on. For a few schools, Fink and Johnstone could not agree, and no rating was assigned for that feature. On one of the features, Improvement (#11), assigning a rating proved very difficult. For example, Fink and Johnstone sometimes felt that a principal had not been in a school long enough to permit credible judgments of the extent to which the school had improved since that principal had come aboard. This variable was eliminated from all further analyses.

Hierarchical cluster analysis was used to examine patterns of intercorrelation among the remaining variables and to construct a reduced set of school descriptor variables that could be used in the regression and path analyses.

## Results and Discussion

### Student Achievement

More than 2,000 fourth-grade students in District 2 took the New Standards examinations in spring 1997. Approximately 92% of them completed the English Language Arts (ELA) examination, and approximately 95% completed the Mathematics examination. Although the district's policy was that all students—including those classified in special education and those identified as LEP—be tested, some students did not complete all three days of testing, a requirement for receiving an official score.

Figure 2 reports the proportion of eligible students for whom ELA scores were actually reported. This figure indicates that in most schools most of the students sat for the ELA exams (median proportion = .96, proportion at the 75th percentile = .98). The exception was a school with 40 fourth-grade students in which the proportion was .63, meaning that 15 of the 40 students did not have ELA scores reported. The proportions for Mathematics are reported in Figure 3 and were slightly higher (median proportion = .97, proportion at the 75th percentile = .99). The .56 proportion tested for Mathematics in Figure 3 was associated with the same school that produced a proportion tested of .63 for ELA.

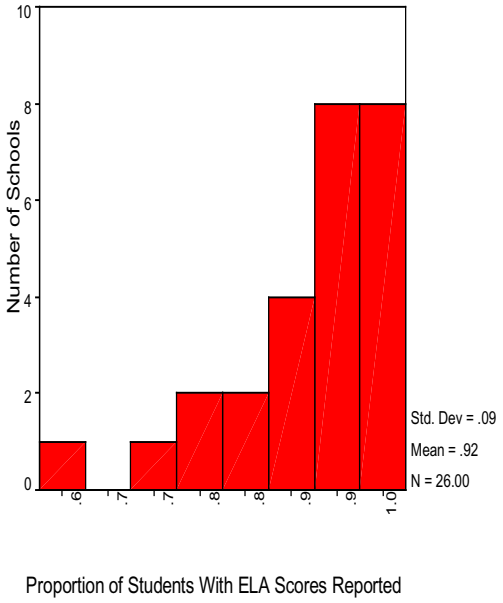


*Figure 2.* Proportion of tested students for English Language Arts (ELA).

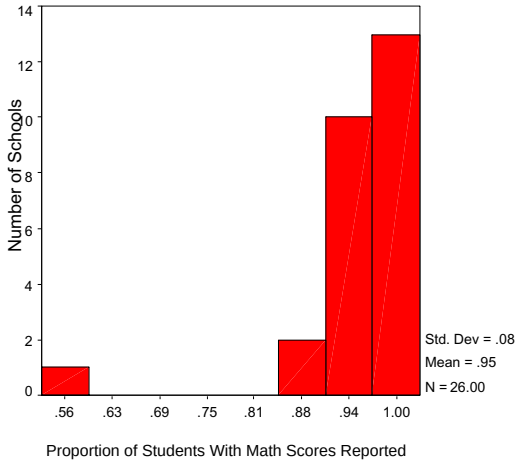


*Figure 3.* Proportion of tested students for Mathematics.

We also investigated whether lower average proportions of reported scores were more likely to be centered in lower SES schools. We found that for Math there was no relation between the proportion of students tested and SES. For ELA, however, there was a significant (quadratic) relationship between proportion reported and SES ($r$ = .56). This relationship remained statistically significant even after omitting the data for the school in which only 63% of the ELA scores were reported and recomputing the correlation. Thus, lower SES schools were more likely to have fewer students complete the ELA examinations than were higher SES schools. Unfortunately, the effect of missing New Standards data cannot be assessed in more detail because the data are at the school level, and, therefore, information about individual students is not available. We will comment later on the possible impact on our findings of lower rates of exam completion in low SES schools.

Figures 4 and 5 show the percentage of fourth-grade students in the district as a whole who met or exceeded the standard in each standards cluster. The percentages of students who met or exceeded the standard in English Language Arts were high compared with other districts with similar demographic characteristics who have also administered the New Standards exams. District 2's performance on the Mathematics standards clusters was not particularly distinguished, however. The difference between Mathematics and English Language Arts achievement reflects the focus of District 2's professional development and accountability efforts over the eight years prior to this study. Until recently, the District had focused most of its professional development efforts on literacy, with only marginal attention given to

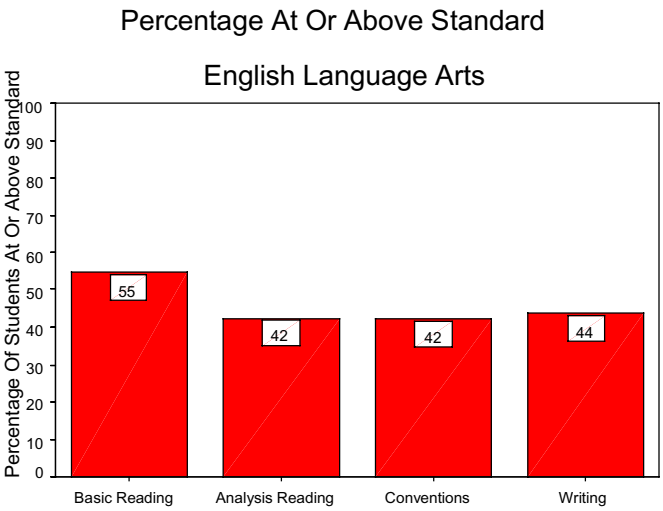


*Figure 4.* Percentage of students at or above standard for English Language Arts.

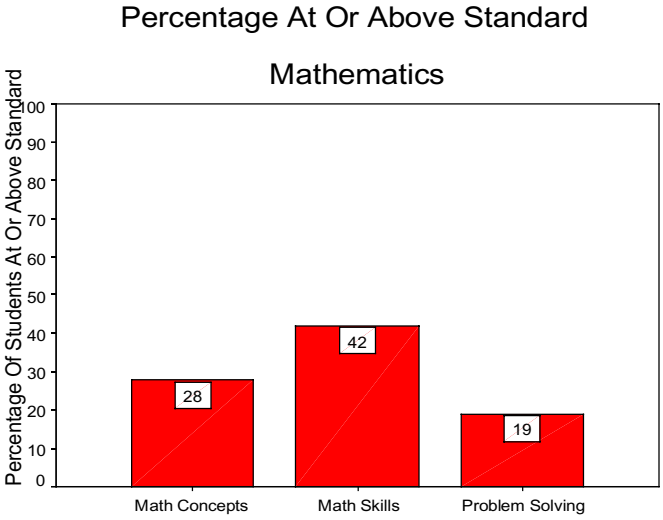


*Figure 5.* Percentage of students at or above standard for Mathematics.

mathematics. This emphasis on literacy may explain the discrepancy in performance between English language arts and mathematics.

District 2 includes a diversity of schools as well as a diversity of students. Some schools have populations almost entirely composed of children from poor families; others have only a few such children. Schools with a predominantly poor population, of course, have a particularly great challenge to meet in attempting to reach high levels of achievement. To evaluate how well District 2 schools are meeting this challenge, we ranked the schools by percentage of students who were not eligible for free or reduced-cost lunches and used these rankings to establish four quartile groupings of schools. These groupings reflect the diversity of students served by District 2, with the low SES schools having more than 90% of their students eligible for free or reduced-cost lunches and the high SES group having less than 21% of their students eligible.

District 2 also has a significant percentage of children classified as LEP. Schools with high proportions of such children, like schools with many poor children, face a challenge in educating students to high standards. We again ranked the schools, this time by percentage of students classified as LEP, and used these rankings to establish four quartile groupings of schools. The low LEP group had less than 6% of their students classified as LEP, whereas the high LEP group of schools had more than 32% of their students classified as LEP.

We then plotted the mean percentage of students who scored above the standard and the mean percentage who scored below the standard by SES and by LEP for each New Standards cluster. Figures 6 through 9 illustrate these plots. (The full set of plots is presented in Harwell & Resnick, 1998.) Figure 6 shows the mean percentage of elementary school students at or above the standard and the mean percentage of students well below the standard for the cluster Reading: Basic Understanding for the four SES groups. As might be expected there is an association between achievement levels and SES. Since the lowest SES schools also had the largest number of unreported scores, the association may be even a bit stronger than the graph shows. The relation between Basic Understanding and LEP (Figure 7) is similar to that for SES.

The relations between achievement in the Math Skills cluster and SES and LEP are plotted in Figures 8 and 9. As shown in Figure 8, high SES students did well (nearly 70% met the standard on skills), but there was a sharp drop for the other quartiles, with especially low performances for the bottom two quartiles. A similar pattern emerged for LEP (Figure 9).

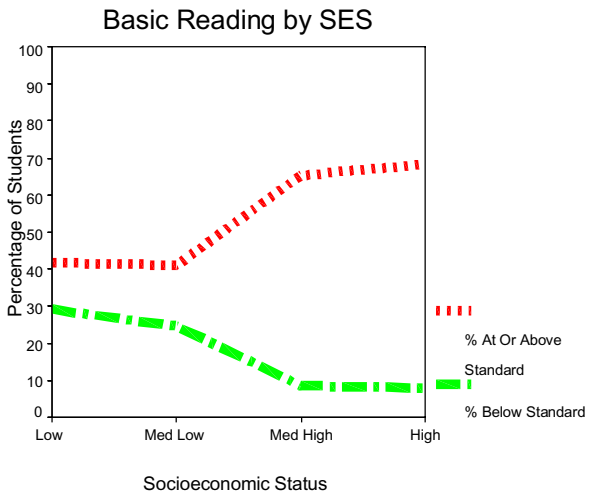


*Figure 6.* Basic reading by socioeconomic status.

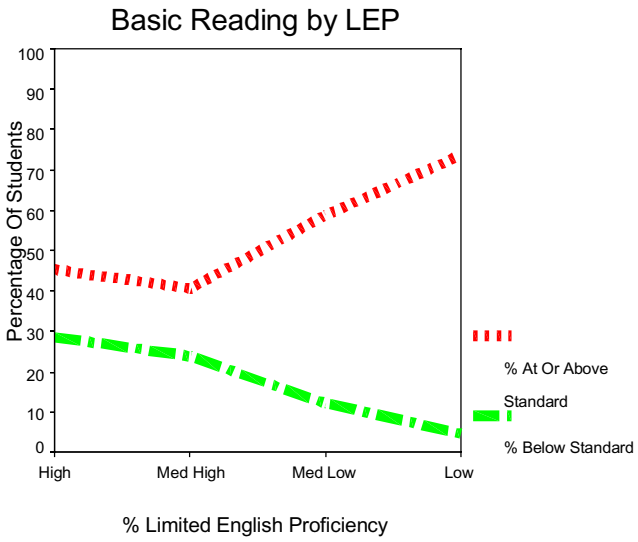


*Figure 7.* Basic reading by limited English proficiency.

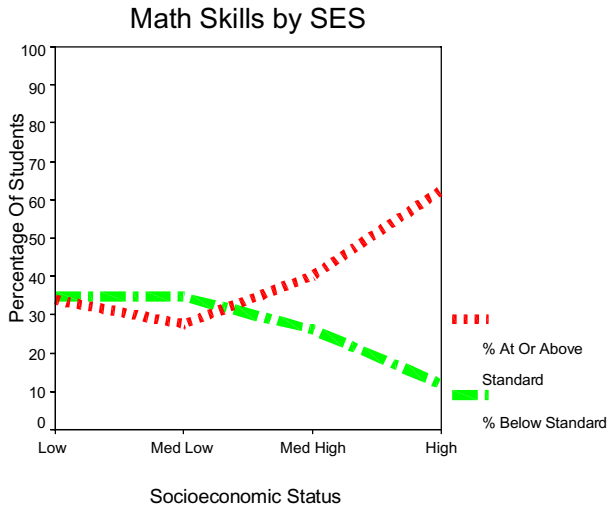


*Figure 8.* Math skills by socioeconomic status.

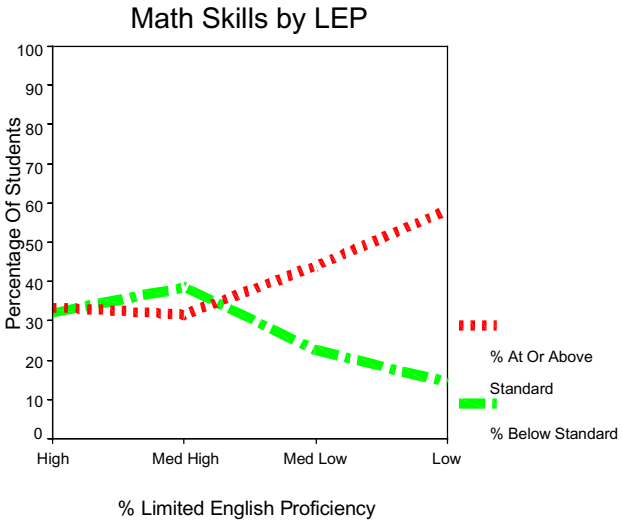


*Figure 9.* Math skills by limited English proficiency.

## Clustering the Rating Variables

Table 2 shows the intercorrelations among the 12 rating variables used in this study. These correlations ranged from a low of .41 to a high of .99, with more than half exceeding .70. In order to explore underlying patterns among these correlations, we performed a hierarchical cluster analysis using the SPSS for Windows (SPSS, 1997) program.

The cluster analysis began by treating the 12 rating variables as separate clusters and then trying to combine clusters until only one was left. At each step, the clustering program identified the cluster of variables with the strongest associations. At step 2, for example, the Global judgment (#12) of the overall quality of the school and Quality of the Staff (#3) were joined. These two variables are correlated at .99, so they are essentially identical judgments. The Global and Quality of Staff cluster was joined in step 4 by the Quality of Teaching (#4) and Quality of Student Work (#1) variables. These two variables represent District 2 leadership's core definition of how a high-quality staff will display its quality: It will do good teaching that will result in very good student work.

Another distinct cluster that was identified about halfway through the clustering was composed of variables that center on the character of professional leadership and professional development in the school—Professional Development (#4), Leadership–Culture (#6), and Leadership–Content (#7). The District 2 theory on the capacity of the school leader to discern good teaching and to select good staff did not emerge until near the end of the clustering and was defined by two variables: Leadership–Discriminate (#8) and Leadership–Select (#9). The two remaining variables, Parents and Community (#5) and Leadership–Weed (#10) did not cluster until the very end, indicating that these features have little in common with the others or with each other. In sum, cluster analyses identified three major clusters of school quality variables—Teaching Quality, Professional Development, and Quality of Staff—that seemed to accord well with the District 2 theory of action.

Next, principal components analysis was used to determine whether the variables comprising each cluster could be aggregated to produce a single variable that could be correlated to student achievement. The principal components results for the first cluster of school quality variables (Global School Quality, Quality of Staff, Quality of Student Work, and Quality of Teaching) strongly suggested that one component underlies these four indices and that each rating variable contributed

Table 2

Correlations Among Engagement Variables for Elementary Schools

| | Quality Student Work | Quality Teaching | Quality Staff | Profes-sional Develop-ment | Parents and Community | Leader-ship Culture | Leader-ship Content | Leader-ship Discrimi-nate | Leader-ship Select | Leader-ship Weed | Global | Potential |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality Student Work | 1.00 | | | | | | | | | | | |
| Quality Teaching | .95 | 1.00 | | | | | | | | | | |
| Quality Staff | .93 | .93 | 1.00 | | | | | | | | | |
| Professional Development | .83 | .87 | .92 | 1.00 | | | | | | | | |
| Parents and Community | .74 | .69 | .66 | .52 | 1.00 | | | | | | | |
| Leadership Culture | .79 | .85 | .85 | .83 | .70 | 1.00 | | | | | | |
| Leadership Content | .86 | .86 | .93 | .93 | .62 | .88 | 1.00 | | | | | |
| Leadership Discriminate | .71 | .64 | .83 | .71 | .49 | .70 | .84 | 1.00 | | | | |
| Leadership select | .84 | .80 | .91 | .82 | — | .72 | .87 | .88 | 1.00 | | | |
| Leadership Weed | .59 | .57 | .67 | .66 | — | .41 | .79 | .46 | .55 | 1.00 | | |
| Global | .94 | .94 | .99 | .90 | .66 | .85 | .92 | .76 | .87 | .61 | 1.00 | |
| Potential | .89 | .88 | .94 | .87 | .67 | .88 | .95 | .85 | .87 | .58 | .95 | 1.00 |

*Note.* The sample sizes used in computing the correlations varied from 17 to 26 because of missing ratings for some indices on some schools. All correlations given are significant at the .05 level.

approximately equally to a Quality factor. We therefore computed a new variable, Teaching Quality, that consists of the simple average of the ratings on the four separate indices. Similar results emerged for the principal components analysis of the Professional Development, Leadership–Culture, and Leadership–Content rating variables, which were then averaged to create another new variable, Professional Development. The Leadership–Discriminate and Leadership–Select variables, which had a high correlation ($r$ = .88), were averaged to create a variable called Quality of Staff. The remaining rating variables, Parents and Community and Leadership–Weed, were treated separately.

**Relation of the Rating Clusters to Student Achievement**

Correlations of the aggregated school rating and achievement variables with each other and with LEP and SES are reported in Table 3. The high correlations between the achievement variables and Teaching Quality and Quality of Staff provide initial evidence that District 2's focus on quality of teaching in schools is crucial to the achievement of students. However, the variable considered central in the District 2 model, Professional Development, had no significant correlation with student achievement. Likewise, the ability of the principal to weed out weak teachers was not significantly related to achievement. On the other hand, the Parents and Community variable, which is not part of the District 2 model, did correlate with achievement. If we had to work from these correlations alone, the hypothesized direct relation between Professional Development and Achievement in the District 2 model (in Figure 1) would have to be rejected, and Parents and Community would have to be added to the model.

We return to the question of Parents and Community later in this report. First, however, we want to examine the relations among the rating clusters initially predicted to influence achievement while controlling for the student variables of SES and LEP. For this purpose, we conducted a series of multiple regression analyses using EnglishAbove, EnglishBelow, MathAbove, and MathBelow as dependent variables and using SES, LEP, Professional Development, Teaching Quality, and Quality of Staff as predictors. In each analysis, SES and LEP were entered into the regression model first, followed by the other predictors. The pattern of results was the same for the four analyses (EnglishAbove, EnglishBelow, MathAbove, and MathBelow), so we report here only on EnglishAbove (see comment 2 in the Technical Appendix).

Table 3

Correlations Among Aggregated School Rating and Achievement Variables

| | English-Above | English-Below | Math-Above | Math-Below | Teaching Quality | Professional Development | Staff Selection | Parents and Community | Leadership Weed | SES | LEP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| English-Above | 1.00 | | | | | | | | | | |
| English-Below | -.94* | 1.00 | | | | | | | | | |
| Math-Above | .89* | -.79* | 1.00 | | | | | | | | |
| Math-Below | -.74* | .61* | -.91* | 1.00 | | | | | | | |
| Teaching Quality | .61* | -.59* | .54* | -.47* | 1.00 | | | | | | |
| Professional Development | .27 | -.29 | .21 | -.16 | .95* | 1.00 | | | | | |
| Staff Selection | .64* | -.65* | .57* | -.52* | .85* | .82* | 1.00 | | | | |
| Parents and Community | .52* | -.53* | .53* | -.40* | .70* | .65* | .48* | 1.00 | | | |
| Leadership Weed | .20 | -.17 | .18 | -.23 | .62* | .70* | .47* | .19 | 1.00 | | |
| SES | .78* | -.84* | .64* | -.54* | .49* | .34 | .70* | .27 | .40 | 1.00 | |
| LEP | -.59* | .63* | -.40* | .32 | -.32 | -.28 | -.47* | -.13 | -.44* | -.73* | 1.00 |

*Note.* The sample sizes used in computing the correlations varied from 17 to 26 because of missing data.

Significant correlations at the .05 level are indicated by *.

Figure 10 shows the standardized slopes that were statistically significant at the .05 level. For example, the slope for SES of .51 means that, with the effects of the other predictors held constant, each single standard deviation increase in the percentage of students in a school who were not eligible for free or reduced-cost lunches is associated with a .51 standard deviation increase in the EnglishAbove variable. This translates to about a 9% increase in the percentage of students at or above standard for EnglishAbove for each 1% increase in the number of students not eligible for free or reduced-cost lunches. The slope of .82 for Teaching Quality indicates a strong relation between this variable and EnglishAbove. The slope for Professional Development was *negative* (-.82), however, *exactly contrary to the District 2 hypothesis.* The predictors LEP and Quality of Staff do not appear in Figure 10 because their slopes were not statistically significant. The failure of Staff Selection, a key component of the District 2 model, to account for any variance in EnglishAbove is especially noteworthy. This pattern held even when the Parents and Community and Leadership-Weed school quality variables (neither of which themselves accounted for a significant amount of variance) were added to the regression.[4]

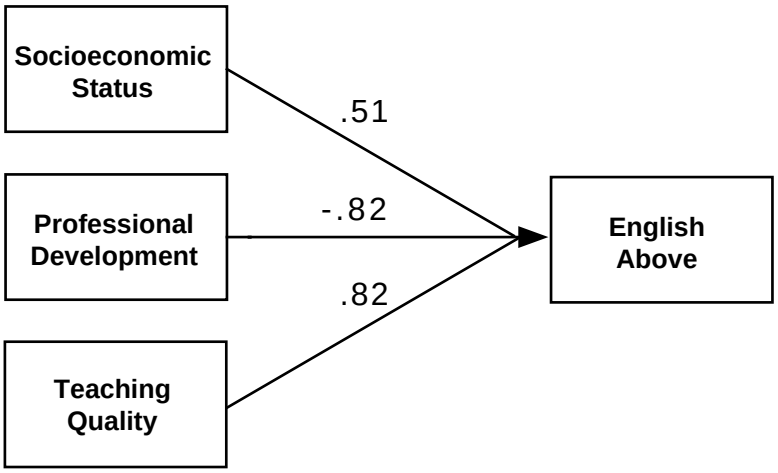


*Figure 10.* Multiple regression analysis of District 2 theory of action for EnglishAbove.

[4] The failure of this analysis and some others reported here to confirm every aspect of the District 2 theory helps to validate our decision to use ratings by highly knowledgeable insiders rather than more "objective" outside observers. Had the individuals doing the ratings been trying to confirm their theory rather than provide the most thoughtful ratings they could, they would probably have given judgments on each feature that were systematically correlated with past, known performances on standardized tests.

The analyses so far appear to call the District 2 model into question. Staff selection has no significant effect on student achievement when all the variables are included in a regression analysis. And professional development, a cornerstone of the District 2 theory, is negatively associated with achievement. However, a more sensitive analysis can be derived using path analysis techniques. As shown in Figure 1, the District 2 model includes the expectation that professional development and quality of staff will influence student achievement through their effects on the quality of teaching in the school. To trace such mediated effects, we conducted a series of separate path analyses for EnglishAbove, EnglishBelow, MathAbove, and MathBelow. Because the results were similar, we report only the EnglishAbove analyses here (see comment 3 in the Technical Appendix).

The results of the path analysis for EnglishAbove are shown in Figure 11. The numbers on the arrows (statistically significant at the .05 level) are the standardized path coefficients reflecting the estimated effect of one variable on another expressed in standard deviations (the numbers in parentheses are the standard errors of the path coefficients).

Comparing Figure 11 with the theory displayed in Figure 1, we see a confirmation of predicted relations between Quality of Staff, Teaching Quality and achievement. There is a similar confirmation of the indirect effect of Professional

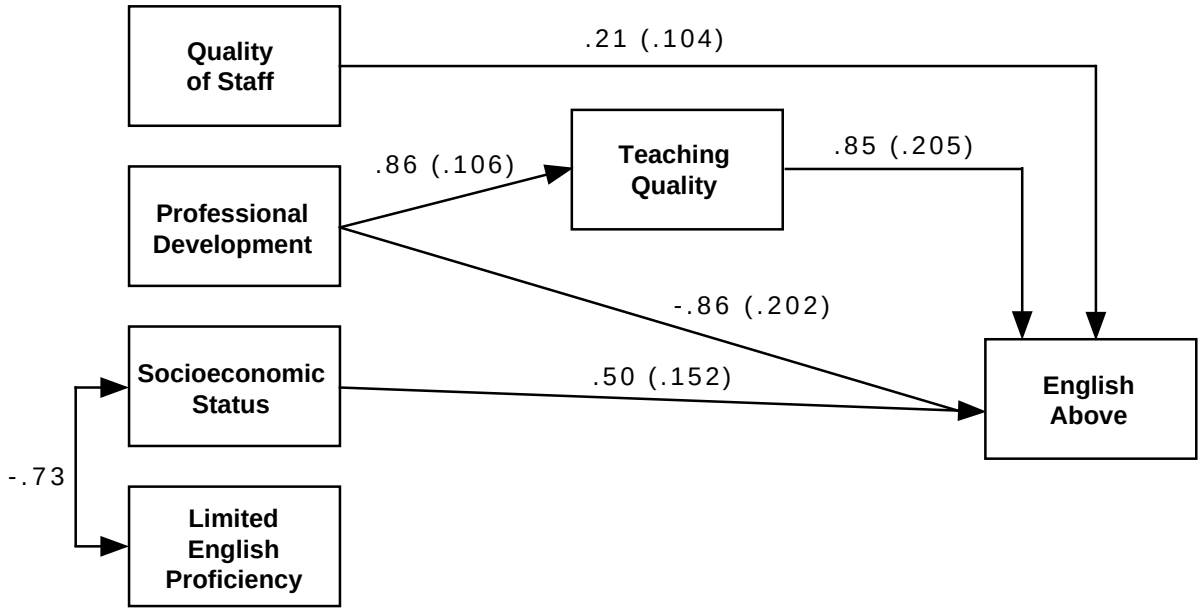


*Figure 11.* Path analysis of District 2 theory of action.

Development on achievement through Teaching Quality. There is again, however, a negative direct association of Professional Development with achievement. Quality of Staff had a relatively weak direct effect on EnglishAbove (.21) but showed no relationship with Teaching Quality. The predicted effect of School Leadership on Quality of Staff and Professional Development was not confirmed and these links do not appear at all in Figure 11. SES had a direct effect on achievement, but did not affect Teaching Quality. LEP did not statistically influence either achievement or Teaching Quality, although its association with SES confirms an indirect (negative) effect on achievement.

In accordance with standard practice, we attempted to clarify the relations among the variables in our path model by rerunning the path model in Figure 11 after omitting the nonsignificant paths. Figure 12 shows the results: There is now a clear pattern in which Professional Development has a strong effect on Teaching Quality, which in turn has a strong effect on EnglishAbove. These path coefficients are very high. Another important result was the indirect effect of Professional Development on EnglishAbove through Teaching Quality (path coefficient = .60), which is not depicted in the figure.

However, the negative direct effect of Professional Development on student achievement appears again. This is puzzling in many ways and we need to consider possible explanations. One possibility is the existence of a few outlier schools that achieve well for reasons other than investment in professional development. To explore this possibility, we returned first to our basic correlational data.

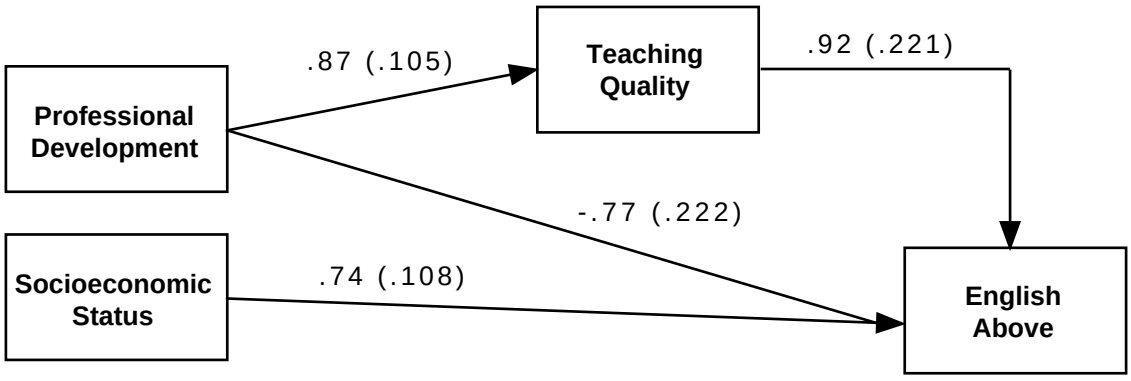


*Figure 12.* Path analysis of simplified District 2 theory of action.

**Off-the-Screen Schools**

Figure 13 shows a scattergram of the correlation between Professional Development and EnglishAbove. Each point is a school. Remember (from Table 3) that the correlation between Professional Development and EnglishAbove was low and not statistically significant. The scattergram shows why. Four schools (those circled at the left of the diagram) are exceptions to an overall pattern of positive correlation between Professional Development ratings and EnglishAbove scores. These four schools show a Professional Development rating that is below the mean on Professional Development ratings (mean = 5.6, median = 6.3) but above average student achievement (mean = 44.3% scoring at or above standard, median = 46.3%).

When these four schools were eliminated from the analysis, the simple correlation between Professional Development and EnglishAbove, which was not statistically significant for the sample of 26 elementary schools, increased to .60 and was statistically significant. What is more, when the path model pictured in Figure 12 was run omitting these four schools, the negative direct relation between

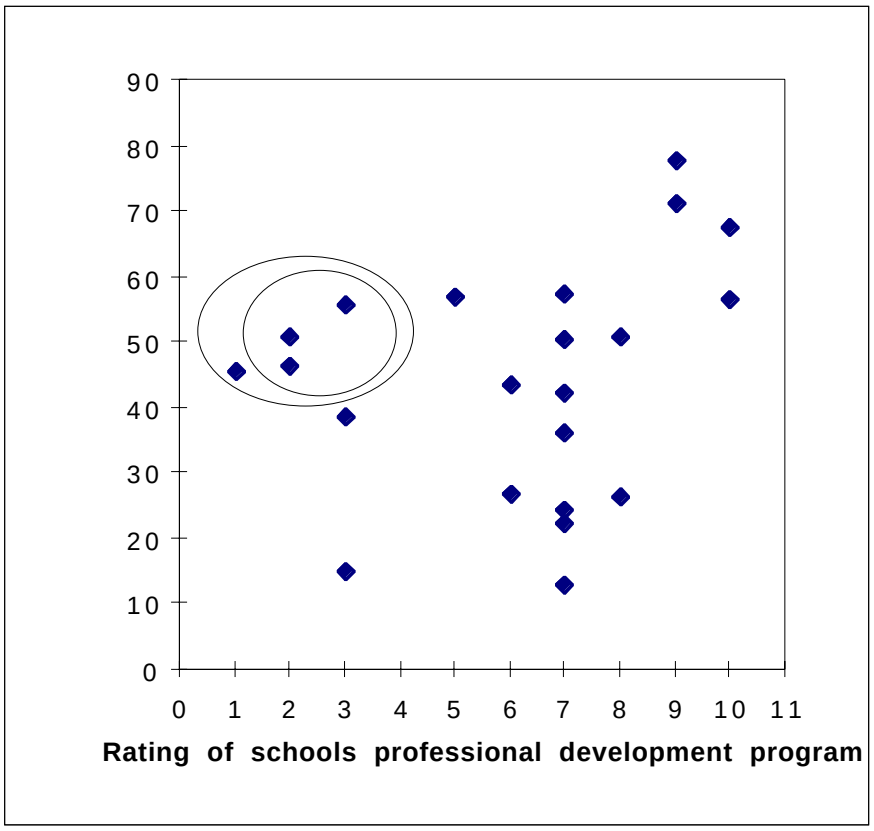


*Figure 13.* Scattergram of Professional Development and EnglishAbove.

Professional Development and EnglishAbove disappeared. The results are shown in Figure 14. Not shown in Figure 14 is a particularly strong indirect effect of Professional Development on EnglishAbove through Teaching Quality (.71).

These are nice statistical results, but unless there is an independent reason for treating the four outlier schools differently from the other schools, it is not legitimate to draw conclusions from them. Are there legitimate reasons to treat these schools differently?

An obvious hypothesis, that all four were high SES schools where intensive professional development was not needed, cannot be sustained. Three of the four outlier schools were in the two lowest quartiles for SES, and two had high proportions of LEP students. A scan of the interviews with Fink and Johnstone that led to their ratings of the schools suggests that they, too, were puzzled about the four outlier schools. They knew that these schools did not "fit" the District 2 theory, but they did not have a coherent alternative theory to explain them.

Fortunately, as we were conducting this statistical study, two colleagues were independently studying District 2 using a qualitative research methodology. Based on interviews with many informants in the district, Elmore and Burney (1997b) identified four classes of schools in the district:

- *With-the-drill:* schools that are models of the District 2 core theory of professional development and teaching quality;
- *Free-agents:* schools that do things somewhat differently from the recommended district framework, but are considered to be leading the way in inventing new forms of instruction and are valued for this leadership;

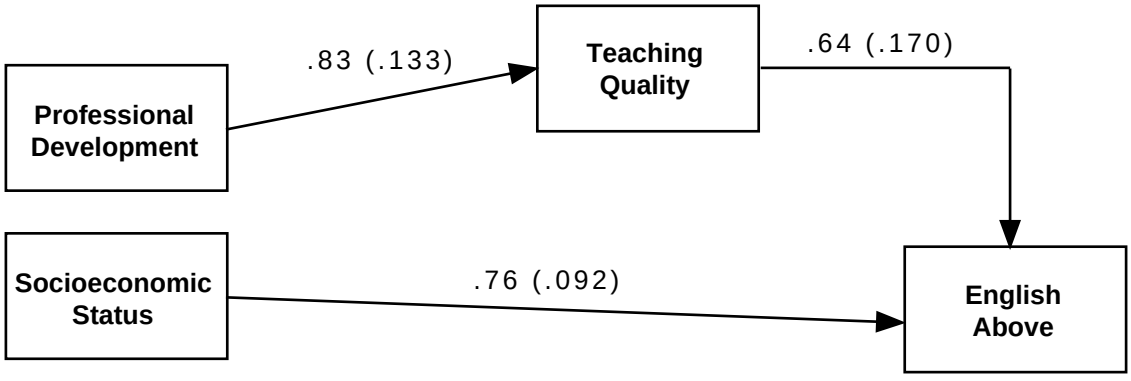


*Figure 14.* Path analysis of simplified District 2 theory of action with four schools removed.

- *Watch-list:* schools that are working within the District 2 framework but need further support and development;
- *Off-the-screen:* schools that are not working within the District framework but are doing passably well on student achievement and are rather well regarded by their parent communities.

Off-the-screen schools are schools that senior administrators in District 2 had decided to leave temporarily undisturbed because they were not failing badly and attempts to intervene assertively would likely produce community reactions that would draw energy away from the instruction and learning work the administrators needed to do in the rest of the district. Senior leadership was, in effect, "picking their battles," choosing to avoid a few fights in order to win many others. The four statistical outlier schools in our study are the four off-the-screen elementary schools in Elmore and Burney's study. We thus had an independent reason to treat these four schools differently.

We next attempted to identify what was different about those off-the-screen schools that accounted for their unexpectedly high achievement level given their ratings of professional development. We compared the four circled (off-the-screen) schools in Figure 12 to the remaining 22 elementary schools on several variables. To do this, we created a dichotomous variable in which the four circled schools were coded 1 and the remaining schools were coded 2. The results of these comparisons are reported in the Technical Appendix, and show that the off-the-screen schools are similar to the remaining schools with one notable exception: Off-the-screen schools show significantly less variability in literacy performance than the remaining 22 schools. This may be the result of fairly rigid adherence to a lock-step literacy curriculum in these schools that reduces variation in student performance. If correct, this explanation would help to explain a piece of the puzzle surrounding these schools, but much remains unknown.

**SES and Achievement**

As we have noted earlier the lower test completion rate in low SES schools may have led to an underestimation of the relationship between achievement and SES in our analyses. In each of our regression and path analyses (Figures 10, 11, 12 and 14) the "true" associations between Socioeconomic Status and EnglishAbove may be somewhat higher than the numbers actually shown. This could, in turn, mean that some other associations would be slightly lower. The associations between

Professional Development, Teaching Quality and English Above, however, are very high, so the basic pattern of results would not be disturbed by a small change in the levels of association.

## Conclusion

We have been pursuing a "statistical detective story," using regression and path analysis methods to determine whether results on a standards-referenced performance assessment were systematically related to the quality of instruction and professional development in the various schools in District 2. Our results broadly confirm District 2's theory of school improvement via school-based professional development. With respect to educational equity, another theme of this series of research reports, the important result here is that the effects of professional development on teaching quality and of teaching quality on examination results hold for both high and low SES (and LEP) schools. Thus, although the District 2 strategy does not eliminate differences between schools with different student populations, it does show that, when poor children receive excellent instruction, they can learn much better than is usually expected. Furthermore, it shows that investment in certain forms of professional development within schools serving poor children is very likely to improve teaching quality in ways that matter for student results.

This study failed to confirm District 2's theory of the role of principal leadership in creating the staff quality and professional development that are so important to quality of teaching in a school. Principal leadership is widely held by other theorists, as well as District 2 leadership, to be a key to effective schooling. We think that our measures of leadership may be more at fault here than the theory. We used five separate measures of leadership, perhaps so many as to induce unreliability of ratings as raters tried to distinguish among dimensions of leadership. In any case, the role of school leadership in creating instructional quality warrants further investigation by those interested in the ways in which standards-based education systems can promote higher achievement and greater equity in schools.

## References


Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum Associates.

Elmore, R. F., & Burney, D. (1997a). *Investing in teacher learning: Staff development and instructional improvement in Community School District#2, New York City*. New York: National Commission on Teaching and America's Future/Consortium for Policy Research in Education.

Elmore, R. F., & Burney, D. (1997b). *School variation and systemic instructional improvement in Community School District #2, New York City*. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, High Performance Learning Communities Project.

Harcourt Educational Measurement. (1997, 1998, 1999). *New Standards Reference Examinations*. San Antonio, TX: Harcourt Inc.

Harwell, M. R., & Resnick, L. B. (1998). *Professional development and teaching quality in a standards referenced education system* (Report to the U. S. Department of Education, Office of Educational Research and Improvement). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Baltimore, MD: The Johns Hopkins University Press.

Howard, J. (1995). You can't get there from here: The need for a new logic in education reform. *Daedalus, 124*, 85-92.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage Publications.

Joreskog, K. L., & Sorbom, D. (1989). *LISREL 7: A guide to the program and applications*. Chicago, IL: SPSS Inc.

Little, R. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Olsen, M. K., & Schafer, J. L. (1998, August). *Models for semicontinuous longitudinal data*. Paper presented at the Joint Statistical Meetings, Dallas, TX.

Resnick, L. B. (1995). From aptitude to effort: A new foundation for our schools. *Daedalus, 124*, 55-62.

Resnick, L. B., & Hall, M. W. (1998). Learning organizations for sustainable education reform. *Daedalus, 127*, 89-118.

SPSS for Windows. (1997). *User's manual.* Chicago, IL: SPSS Inc.

Stein, M. K., & D'Amico, L. (1999). Instructional improvement in New York's Community School District #2: The role of administrators' knowledge. In J. P. Spillane, (chair), *Policy implementation and cognition: school leaders and the implementation of instructional policy*. Symposium presented at the 1999 annual meeting of the American Educational Research Association, Montreal.

Stein, M. K., & D'Amico, L. (1998). *Content-driven instructional reform in Community School District #2.* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, High Performance Learning Communities Project.

Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn, and equity: New Standards examinations for the California Mathematics Renaissance* (Tech. Rep.). Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

## Technical Appendix

### Comments

1. Both unweighted data values (percentages), which do not take into account varying numbers of students across schools, and weighted values, in which the percentages for each school were weighted by that school's contribution toward the total, across-school sample size, were computed and used in many of the analyses. Results from the unweighted versus weighted cases produced quite similar results, probably because the number of students who completed the New Standards examinations at most schools (median = 78) ensured good precision in estimating parameters in the data analysis. As a result, unweighted values were used in the data analyses that are reported.

2. Unfortunately, there were missing data for some of the predictor variables, a problem that was exacerbated by the relatively small number of schools. Because different methods for handling missing data can produce different statistical results, we ran the regression analyses for the data produced by each of three methods of handling missing data; if all three methods produced similar results, the dependency of the findings on the way that missing data were handled would be lessened; if the results varied as a function of how missing data were handled then it would be necessary to choose among the methods.

First, we ran the regression analyses after deleting schools with missing data on any of the predictor variables, which reduced our sample size of elementary schools to 19 (so-called listwise deletion). Second, we used the mean imputation option available in the regression program in SPSS for Windows (1997) in which missing values are estimated by imputing the mean of that variable computed using available data (Little and Rubin, 1987, describe this method and its limitations). Third, we used the AMOS program (SPSS, 1997) to perform the regression analyses because of its ability to estimate regression parameters in the presence of missing data using a procedure described in Arbuckle (1996). This procedure requires that the data be missing at random (MAR), implying, for example, that the reason(s) for a missing rating for a school on an engagement variable could be predicted by available data. We felt that this assumption was reasonable and used AMOS to perform the regression analyses, which used available data from all 26 schools (Olsen and Schafer, 1998, indicate that this may be a reasonable approach even if MAR is not satisfied). On the whole, the three methods produced similar results, although, as expected, mean imputation resulted in more conservative findings, such as slightly smaller $R^2$ statistics and standardized estimates of the regression parameters. We report the slightly more conservative results obtained with mean imputation.

3. Path analyses were done using the AMOS (SPSS, 1997) and LISREL (Joreskog & Sorbom, 1989) computer programs. We chose AMOS because of its ability to estimate path coefficients in the presence of missing data, allowing us to use whatever data were available for the 26 elementary schools, and because it allows the sensitivity of the standard errors associated with tests of the path coefficients to non-normal data to be assessed using a bootstrap procedure (see the AMOS

manual). However, AMOS does not provide tests of indirect effects even if the data are complete (i.e., there are no missing data). LISREL, on the other hand, cannot estimate path coefficients in the presence of missing data but does provide tests of indirect effects for complete data. By using both programs to perform the path analyses we hoped to be able to use as much of the data as possible and still be able to test for an indirect effect of Professional Development on achievement through Teaching Quality.

Path analyses require a careful evaluation of model-data fit before the results can be credibly interpreted (Hu & Bentler, 1995). Our strategy was to use AMOS to estimate path coefficients in the presence of missing data, and then use LISREL to estimate path coefficients for the model in Figure 10 using schools with no missing data on the variables in this figure. If the two sets of path coefficients were similar we could assess model-data fit using the results for the schools with no missing data ($N$ = 22) but report the path coefficients based on all 26 schools; if the two sets of coefficients disagreed, it would be necessary to delete the schools showing missing data and base all of our analyses and interpretations on the reduced sample. Fortunately, the two sets of coefficients were quite similar, and we treated the two samples as interchangeable and report the coefficients produced by AMOS. Model-data fit was assessed using the chi-square goodness-of-fit test, the GFI, NFI and CFI fit indices, and the standardized model residuals. For every path model we examined except one, the chi-square fit test was not significant at the .05 level of significance. However, the three fit indices always exceeded .90 as recommended by Hu and Bentler (1995), and almost all of the standardized model residuals were less than plus or minus two as recommended by Hayduk (1987). There was also no evidence that the assumption of normality was seriously violated.

4. The off-the-screen schools were compared to the remaining schools on several variables for means, variances, and identity of distributions (see Table A1 below). None of the tests of two independent means were statistically significant using the Welch-Aspin (separate variance) test except for Professional Development and Teaching Quality, which is not surprising because these are important criteria used in categorizing a school as off-the-screen. None of the tests of identity of distributions were significant. Although tests of variances were not significant for Math, large variance differences for ELA appeared, with the off-the-screen schools showing less variation than the remaining schools. In fact, the ratios of the variance of the off-the-screen schools to the others ranged from 7:1 to 19:1. This indicates homogeneity of performance among the off-the-screen schools that did not exist among the remaining schools.

Table A1

Comparison of Off-The-Screen and Remaining Schools

| Variables | Hypothesis 1 $\sigma_1^2 = \sigma_2^2$ | Hypothesis 2 $\mu_1 = \mu_2$ | Identity of distributions |
|---|---|---|---|
| Reading Basic | Sig. | Not Sig. | Not Sig. |
| Reading analysis | Sig. | Not Sig. | Not Sig. |
| Writing | Sig. | Not Sig. | Not Sig. |
| Conventions | Sig. | Not Sig. | Not Sig. |
| Math Concepts | Not Sig. | Not Sig. | Not Sig. |
| Math Skills | Not Sig. | Not Sig. | Not Sig. |
| Math Problem Solving | Not Sig. | Not Sig. | Not Sig. |
| EnglishAbove | Not Sig. | Not Sig. | Not Sig. |
| MathAbove | Not Sig. | Not Sig. | Not Sig. |
| Proportion Reported | Not Sig. | Not Sig. | Not Sig. |
| Professional Development | Not Sig. | Sig. | |
| Teaching Quality | Not Sig. | Sig. | |

*Note.* Statistical tests performed using $\alpha = .10$. Test of $\sigma_1^2 = \sigma_2^2$ done using the Levine test, tests of $\mu_1 = \mu_2$ done using Welch-Aspin *t* tests.