Pathway to Results:

Pay for Performance in Denver

December 2001

ctac | COMMUNITY TRAINING AND ASSISTANCE CENTER

ABOUT CTAC:

The Community Training and Assistance Center is a national not-for-profit organization with a twenty-two year track record of success in urban communities. It focuses on developing leadership, planning and managerial expertise within community-based organizations, school systems, collaborative partnerships, state and municipal governments, and health and human service agencies. The Center has provided assistance to hundreds of community-based organizations, coalitions and public institutions in the United States and several foreign countries.

The Center's staff is comprised of nationally recognized executives, educators, policy makers and organizers who have extensive experience working with city, county and state agencies, educational institutions, federal legislative bodies, not-for-profit organizations, philanthropic institutions and the private sector.

Pathway to Results:

Pay for Performance in Denver

December 2001

COMMUNITY TRAINING
AND ASSISTANCE CENTER

Acknowledgements

This research study is sponsored by the Denver Public Schools Board of Education and the Denver Classroom Teachers Association.

CTAC would like to begin by thanking the many members of the Denver community who took time out of their busy schedules, sometimes at significant inconvenience, to spend time with CTAC staff members discussing issues related to the impact of the Pay for Performance Pilot. We would also like to thank the hundreds of teachers, administrators, parents and other individuals who took the time to fill out and return our surveys. Without the willingness of these individuals to share their views, there would be much less data to analyze and many fewer findings to report.

CTAC would like to acknowledge the many people who have contributed both to the pilot and to making this study possible. The leadership of the Board of Education has been pivotal to the pilot. This has involved all board members: President Elaine Gantz Berman, Sue Edwards, Lucia Guzman, Sharon Macdonald, James Mejía, Bennie Milliner and Les Woodward. The leadership of the Association has also been essential, including President Becky Wissink, former President Andrea Giunta and Executive Director Bruce Dickinson.

We have been assisted by many leaders of the Denver Public Schools, including Superintendent Jerry Wartgow and former superintendents Bernadette Seick, Chip Zullinger, Sharon Johnson and Irv Moskowitz. They have been uniformly helpful.

This study could not have been conducted without the assistance of Design Team Leader Brad Jupp, current Design Team members Cheryl Betz, Henry Roman and Shirley Scott, former members Pat Sandos and Becky Wissink, and current assistant Ginger Doyle. Their energy, enthusiasm and commitment are the guiding lights of the pilot.

We would like to extend a special thank you to the teachers, administrators and parents at the pilot schools. They are making special efforts on behalf of students.

We have been assisted by many current and former members of the central administration. While there are too many such people to mention everyone individually, we would like to identify the following with our appreciation: Assistant Superintendent Wayne Eckerling; Pilot Champion Richard Allen and former champions Larry Lindauer and Craig Cook; Beth Celva, Dianne Lefly, James McIntosh and Richard Machado of Assessment and Testing; Ethan Hemming of Educational Services; Joyce Fell of Human Resources; Director Aaron Gray and Flor Amaro of Community Relations; and Cindy Rundstrom of the Superintendent's Office. They have all been generous with their time and attention.

The Center is grateful for all those listed above, and the many other people who helped in making this study a reality.

This study has been made possible through the generous support of Rose Community Foundation, Donnell-Kay Foundation, The Sturm Family Foundation, The Piton Foundation and an anonymous donor.

The findings, analyses and recommendations expressed in this study are those of the Community Training and Assistance Center.

Credits

This study was conducted and prepared by the Community Training and Assistance Center of Boston, Massachusetts.

Denver Project Team

Denise A. Bell ★
Lee Bray
Peggie L. Brown
Robin C. Burr, Ph.D.
William M. Eglinton
Roberta A. Glass
Donald B. Gratz, Ph.D.★
Barbara J. Helms, Ph.D.★
Mimi Howard
Donald W. Ingwerson. Ed.D.
William C. Lannon
Candy M. Miller
Marcia J. Plumleigh, Ph.D.
William J. Slotnik★
Maribeth Smith ★
Lynn J. Stinnette
Martha Swartz

Statistical Measurement Task Force

Robert H. Meyer, Ph.D.
John B. Willett, Ph.D.

★ *Principal Study Authors*

Contents

Executive Summary

In September 1999, the Denver Public Schools (DPS) and the Denver Classroom Teachers Association (DCTA) jointly initiated the Pay for Performance (PFP) Pilot. This four year pilot is unique. It is a joint union/district collaboration designed to link teacher compensation with student achievement. By the close of the pilot, the parties intend to develop a new salary structure that will be based, in part, on student achievement.

This initiative is being conducted at a set of pilot schools, whose results are compared to a group of control schools. Teachers in pilot schools develop two objectives based on student achievement that must be approved by their principals. Three approaches are being used for objective setting and to determine progress in meeting those objectives. Teachers receive additional compensation if they meet their objectives. The pilot is being stewarded by a four person Design Team, comprised equally of district and union appointees and now reporting directly to the Superintendent.

The Community Training and Assistance Center (CTAC) was commissioned in November 1999 to conduct a comprehensive study of the impact of the pilot, and to provide ongoing technical assistance to help assure a pilot of quality and integrity. The study has two formal reporting points. The first, at the pilot's midpoint, is presented in this volume. The second will be presented in November 2003.

The study has four central components. First, it examines the impact of the pilot on student achievement by comparing three measures of student achievement across the pilot and control schools. Second, it closely examines teacher objectives: their substance, quality and the extent to which they correlate with achievement. Third, the study examines a range of school, teacher and student factors which impact teaching and learning and may have affected the pilot. Fourth, it considers the broader institutional factors that may have affected implementation.

This report is formative. It covers the baseline year of the pilot (1999-2000) and the subsequent full year of implementation (2000-2001). The report analyzes results as seen in student achievement data and as perceived by teachers, parents and administrators. In so doing, the study examines the impact and activity of the pilot in detail, identifies issues that will need to be explored further and addressed, and recommends steps to enhance the overall success of the pilot.

The report presents findings based on the halfway point of the pilot. The pilot is very much in the critical phase of seeking to fully and fairly test a powerful concept. The question is whether pay for performance is a viable and effective strategy for the Board of Education and the Association to use to accomplish their goals. The full four years of the pilot will be needed to provide sufficient evidence to answer this question.

In this study, CTAC conducted and reviewed more than 300 interviews, examined survey responses from more than 480 teachers, parents, administrators and others involved in the pilot, analyzed thousands of student records plus teacher and school demographic data, and reviewed hundreds of documents pertinent to the pilot's operation.

This report contains CTAC's analyses, findings and mid-point recommendations. These are highlighted below. The issues are complex and multi-faceted, and are discussed in full detail in the chapters of the report.

A. Primary Findings

Impact on Student Achievement

- On the Iowa Test of Basic Skills (ITBS) from Spring 2000 to Spring 2001, the pilot elementary schools decreased slightly more in math than the control schools. The control elementary schools gained slightly more in reading and language.
- The pilot middle school out-performed ten of the seventeen control middle schools in student gain on the ITBS reading test and out-performed fifteen of the seventeen control middle schools on the ITBS language test.
- Colorado Student Assessment Program (CSAP) total reading scale scores are higher for the third and fourth grade students at the pilot schools than the control schools for 1999-2000 and 2000-2001. Due to the current structure of the CSAP, longitudinal comparisons of individual student performance from one year to the next cannot be made.
- Students whose teachers developed the highest quality objectives, based on a rubric developed by CTAC, average greater gains in achievement on the ITBS—whether these objectives were met or not met—than students whose teachers' objectives were scored lower on the rubric.
- Similarly, students whose teachers developed the highest quality objectives also scored higher on average on the CSAP—whether the teachers met their objectives or not for compensation purposes.

Impact of Objectives

- Teachers see the objectives as a significant element of the pilot—the connector between student achievement and teacher compensation.
- The objectives written in the first two years of the pilot were evaluated for their rigor and overall quality, using a four-point rubric. When teachers had the highest quality objectives (rubric score 4), their students showed increases in achievement on both the CSAP and the ITBS regardless of whether the objectives were met or not met.
- The quality of teacher objectives (rubric score 4) correlates positively with student gain, regardless of the pilot approach used at the school.
- There are similarities between the objectives written at the pilot schools and control schools. However, the use of baseline data is more evident at the pilot schools.
- There is only a minimal relationship between the teacher objectives and the school plans.

Perceptions of Teachers, Administrators and Parents

- While many respondents were initially concerned about an increase in competition among teachers, results to date suggest that competition has stayed the same or decreased slightly at most schools. Many teachers also see an increase in cooperation among teachers.
- Many teachers believe that quality teaching and positive results should be rewarded. However, teachers are skeptical of some of the concepts of Pay for Performance.

- Most teacher groups believe that fair objectives can be set, with specialists having the least confidence that fair objectives can be set. There are concerns, though, about the reliance on single measures.
- A majority of teachers indicate that they are not doing anything differently. At the same time, nearly half also say that PFP has led to a greater focus on student achievement at their schools.
- Parents also believe that the pilot has increased the focus on student achievement.
- More than 70% of the pilot school administrators believe that, by the end of the pilot's second year, student achievement increased. Nearly half of the teachers felt this way.
- Participants value the training they have received, but express a significant need for more professional development.

Institutional Factors

- The tasks of producing individual student achievement data for teachers, and of linking student results to specific teachers, has proven more difficult than originally anticipated.
- The highly charged nature of accountability within the state and nation has adversely affected the climate for implementing the pilot.
- While considerable progress has been made, difficulties of measurement and assessment remain significant. In the final two years of the pilot, further steps are needed to assure that assessments are available that fairly and accurately measure results for all teachers.
- The inclusion of special subject teachers, specialists, and special education teachers, as well as the inclusion of secondary schools within the pilot, add to the difficulties in both classroom assessment and data management.
- The Design Team has been a catalyst for progress in the pilot.
- The district's awareness of the importance of using student achievement data to improve instruction has grown considerably due to the pilot.
- The ability of the Board of Education and the Denver Classroom Teachers Association to make mid-course corrections as necessary has been a critical strength of the pilot.

B. Recommendations

The primary recommendations are presented in Chapter X. These recommendations are based on findings presented throughout the report, and on the summary of critical issues presented in Chapter IX. The report makes recommendations in the following areas:

Issue One: Objectives

The objectives are the linchpin of the pilot. The study has found that objectives that are of high quality, and specify learning content, are positively correlated with greater gains in student achievement, regardless of whether the teacher actually met the objective. Other findings concern the difficulty many teachers have in writing objectives, especially the many teachers who teach special subjects, special education, or who are in non-instructional areas such as psychologists and nurses.

Recommendations include:

- Emphasize learning content in future objectives.
- Provide teachers with more support and options in objective writing.

- Address the fairness concerns identified by special subject teachers, specialists and special education teachers.
- Establish a direct relationship between individual teacher objectives and each school's improvement plan.

Issue Two: District Data Capacity

A teacher performance system that is based, in part, on student achievement must start with the ability to directly link information on teachers, students and classes. Despite considerable technical proficiency within the district, Denver's departmental data systems have historically been separate from each other and were not designed for the needs emerging from the pilot. Linking teacher and student data systemwide, as well as creating more sophisticated means of reporting results, requires a cooperative, inter-departmental effort.

Recommendations include:

- Build an integrated data system. Convene key departmental representatives to recommend how linkages between different departments, and critical issues such as the lack of consistent teacher identification, should be addressed. This includes establishing a mechanism to identify and resolve inter-departmental data support issues.
- Establish and implement uniform expectations and standards for administering major district tests.
- Broaden district capacity to provide teachers with appropriate data on student learning through the further development of the OASIS intranet system.

Issue Three: Assessment

The assessments upon which objectives are based and judgments of success are made must be fair, valid and appropriate. Further, the set of assessments being used must be specific enough to reflect the instructional content of a particular class, and broad enough to provide a fair measurement across classes and schools. No single current assessment meets all of these purposes.

Recommendations include:

- Develop a means for using multiple measures at the classroom level.
- Select and align assessments by grade, level and subject, so they reflect what teachers have taught and students are intended to learn.

Issue Four: Professional Development

The writing of objectives based on increased understanding of student achievement data, which serves as a basis for the pilot, requires teachers and principals to learn new skills. While professional development to date has advanced this knowledge, more is needed. Further, the district's professional development efforts are not consistently applied or evaluated, and may not be aligned with the most critical needs at the schools.

Recommendations include:

- Provide expanded professional development to teachers and principals at the pilot sites.
- Establish department leaders and an appropriate structure to plan, coordinate and provide professional development activities at the district level.

- Conduct an audit of professional development activities.
- Begin to prepare for the post-pilot transition. As the learnings from the pilot move to scale, many of the Design Team's leadership functions will need to be filled by central departments.

Issue Five: The Pilot and a New Salary Structure

The pilot's statement of purpose sets forth a goal of using student achievement as part of a new compensation system. As the district consolidates pilot learnings and explores options for that new system, a key task will be to consider the implications of taking any new system to scale. A joint task force has been impaneled to research possibilities and recommend options for consideration by the Board of Education and the Association.

Recommendations include:

- Reposition the pilot in relation to a new teacher salary structure. In this context, and in response to the fairness issues emerging through the pilot, the sponsoring parties might consider the possibility of basing part of the compensation on pay for performance—as determined by the performance of teachers over a multi-year period.

Issue Six: The Approaches

The pilot is constructed around three approaches, two based on differing assessments and one based on a teacher's acquisition of skills and knowledge. The study's findings to date indicate varying results for the three approaches. However, a single approach has not emerged as superior in terms of increasing student achievement. These approaches will continue to be studied, but another option is suggested by the current findings.

Recommendations include:

- Begin a longer term effort to integrate the three approaches. As the pilot proceeds along its current course, steps can be taken to move in the direction of a needed change. The best potential result appears to be a system in which all classroom results are assessed with multiple measures and teachers receive the necessary professional development.
- Convene key departments and units to develop the method and timeframe for integrating the three approaches.

Issue Seven: Organizational and Financial Forecasts

There is a range of costs associated with the implementation of a pay for performance system. These include the financial costs which result from new fiscal outlays. They also include the institutional costs related to changing practices and redirecting existing district resources.

Recommendations include:

- Project the costs of implementation. This includes a two-year projection covering the balance of the pilot and a five-year projection covering the costs of bringing the pilot to scale.
- Include projections for both direct and indirect costs.

C. Summary

The Pay for Performance Pilot has revealed the complexity of creating a system that links teacher compensation, in part, to student achievement. The pilot has also embarked on a series of steps that show promise of achieving that linkage. District, union and pilot leaders are confronting challenges related to fairness, classroom assessments, professional development and the importance of greater organizational alignment. They are demonstrating both commitment and courage by focusing on these continuing challenges in a way that shows considerable promise of success.

The pilot is providing a vehicle for experimentation and field-testing of an extremely powerful concept—pay for performance. Denver is exploring the thorny issues that have always prevented any link between teacher compensation and student achievement, a source of great concern to educators, policy makers and community members. Denver is seeking to develop a system that links compensation and achievement fairly, rewards teachers for quality performance, and provides the broader community with an exemplary, more accountable district. In so doing, Denver is forging a new pathway to results.

CHAPTER I

Overview

A. Background and Charge

The Denver Public Schools and the Denver Classroom Teachers Association are co-sponsoring a four-year initiative in teacher accountability and improvement, the Pay for Performance (PFP) Pilot. This pilot is designed to establish a linkage between teacher compensation and student achievement.

During each year of the initiative, each participating teacher sets two classroom-specific objectives which must be approved by the principal at the school. The teachers are awarded additional compensation if they meet these objectives. Three approaches are being used for objective setting and to determine progress in meeting these objectives. Approach One is based on student progress on a norm-referenced test (Iowa Test of Basic Skills). Approach Two is based on student progress on a criterion-referenced test. Approach Three is based on a teacher's demonstrated acquisition of skills and knowledge, but also requires links to student achievement. In addition to the pilot schools, the district has also designated a group of control schools for the pilot. The entire initiative is being stewarded by the Pay for Performance Design Team, comprised equally of district and union appointees.

By the close of the pilot, DPS and DCTA are intending to develop a new salary structure that will be based, in part, upon student academic achievement. Moreover, they are collaborating to help the district become more focused on increasing student achievement and improving overall professional performance.

The essence of a pilot is experimentation. As such, the sponsoring parties sought to approach the issue of experimentation strategically. They jointly agreed to commission a comprehensive research study to identify the impact of the pilot and to explore the range of factors that contribute to that impact.

There are two key benchmarks for this research study. The first is this interim report, a formative report which delineates results of the pilot, covering the baseline year (1999-2000) and the subsequent full year of implementation (2000-2001). The analyses and findings contained in this interim report provide the necessary foundation for making mid-course adjustments to the pilot. The interim report will be presented initially in December 2001, and will be followed by a final, summative report. The final report will be presented to the sponsoring parties in November 2003.

This pilot and study are being undertaken during a period of intense national focus on accountability as the means to improve public education. In this context, Denver is firmly committed to examining the impact of the pilot on students, teachers, schools and the district. The intent is to guide local practice and decision-making, while concurrently informing state and national policy-makers.

B. Community Training and Assistance Center

In November 1999, following a national search by the Design Team, the Community Training and Assistance Center (CTAC) was selected to conduct the comprehensive study of the impact of the pilot. CTAC was also asked to provide technical assistance to help build district capacity so that a pilot of quality and integrity could be implemented and studied.

CTAC has been a leading provider of technical assistance to school systems throughout the United States since 1979. In particular, CTAC provides multi-year, comprehensive technical assistance to major urban school districts attempting to improve student achievement, community involvement, and overall school performance and accountability. At the community level, CTAC also assists more than 90 nonprofit organizations and community agencies each year.

C. Areas of Inquiry

This research study has four umbrella components. These over-arching components collectively focus on results and the key factors which may contribute to these results.

Impact on Student Achievement

The foundation of the pilot is student achievement. Accordingly, the foundation of the study is to examine the relationship between Pay for Performance and actual results in student achievement. The study examines (a) the changes in student achievement which have occurred at the participating pilot schools, and (b) how these changes in achievement at the pilot schools compare to those of the district-designated control schools.

Impact of Objectives

At the school site level, the objectives set by individual teachers are the centerpiece of Pay for Performance. Teachers receive additional compensation only when they meet their objectives. The study examines both the substance and impact of the objectives. It further examines the correlation between teachers meeting their objectives and actual increases in student achievement.

School, Teacher and Student Factors

There are site level factors which should be considered beyond compensation that may influence student achievement. Pay for Performance exists in a broader school context. Therefore, the study examines a range of factors—school, teacher and student—that may contribute to, and may prove to enhance, the achievement of students or the effectiveness of the pilot, and those factors that are associated with less success for students or teachers.

Broader Institutional Factors

The pilot also exists in a broader district context. Accordingly, this study examines a range of institutional factors that may support or impede the pilot. Pay for Performance is an initiative with significant systemic implications. The study analyzes which efforts are perceived by various constituencies as supporting or impeding the progress of the pilot, which lessons have emerged from the pilot, and the implications of those lessons for the district.

D. Methodology

The study is based on several primary sets of data: detailed results from individual and group interviews, comprehensive surveys of pilot and control schools, substantial data on student achievement, teachers and schools, and an extensive range of internal and external documents. The study design is discussed in detail in Chapter III of this report.

These data have been subjected to several layers of analysis. Statistical tests were conducted on quantitative data to provide statistically valid comparisons between the responses of different groups of people on certain survey and interview questions, for example, and to establish relationships between these responses, groupings of schools and the data on student achievement. Further, based on the availability and condition of the data, CTAC conducted a range of detailed analyses of student achievement and related data, including regression analysis, chi-square analysis, multi-level modeling, hierarchical linear modeling, and value-added modeling.

A brief description of data sources and limitations is contained below, and sources are referenced throughout the text of the report. Quotations from interviews and analysis based on survey data are not specifically referenced, but all are specifically annotated within CTAC's files.

The recommendations are drawn from extensive analyses of these data sources, plus research on school practices, district management and educational measurement, and CTAC's national experience in urban school districts engaged in reform initiatives, performance assessments and issues related to accountability.

Comprehensive Surveys

The first data set was developed from a series of comprehensive surveys, which were prepared and distributed at periodic junctures to stakeholders at both pilot and control schools, including teachers, administrators and parents. The parent surveys were made available in English and Spanish. The purpose of these confidential surveys was to establish perceptual baselines, elicit opinions from the broadest possible range of school stakeholders as to the status of the pilot, and identify the extent of perceived change during the initial two years of the pilot.

Survey instruments for administrators and teachers were distributed and gathered at the school level in sealed envelopes. All pilot school teachers received surveys, as did a random sample of control school teachers. A random sample of both pilot and control school parents also received sealed surveys via mail. All surveys were shipped directly back to an independent scanning subcontractor.

Individual and Group Interviews

The surveys were supplemented by a series of more than 300 confidential interviews, conducted in English and Spanish, both individually and in groups. Interviewees included teachers, administrators and parents at the pilot and control schools, members of the Board of Education and the Association, an extensive range of individuals from the central administration, and from political, philanthropic, media and business sectors.

Student Achievement and Documentary Data

Data on student achievement and teacher objectives were provided by the Denver Public Schools. These data were supplemented by information drawn from the Colorado Department of Education. These are highlighted in the next section and further delineated in Chapters III through VII of this report.

The subject of this study is the impact, at many levels, of a complex and multi-faceted change initiative. CTAC examined in detail documents made available by the sponsoring parties and all relevant support departments. These are referenced throughout this report.

Customization

Each component of the study has a customized methodology:

- The first component focuses on determining gains in student achievement. This involves analyzing student achievement data from the different assessments most commonly used within each of the three pilot approaches and within the district. Specifically, CTAC has analyzed student achievement using such measures as the Iowa Test of Basic Skills (ITBS), the Colorado Student Assessment Program (CSAP), and 6+1 Trait Writing (Six-Trait), and has also reviewed data from the Colorado Basic Literacy Act assessments (CBLA). Where available, student demographic and behavioral data have also been examined.

- The second component focuses on the teacher objectives that have been established at each of the pilot schools. This analysis has been supplemented by comparing the pilot objectives to a sampling of teacher objectives established at the control schools. The teacher objectives at the pilot schools have also been studied in the context of the respective school improvement plans.
- The third component on school, teacher and student factors is based on a range of data sources. These include confidential surveys and interviews which CTAC has conducted regularly. They also include district-provided information, as available, on school programs, school/class size, demographics, principal backgrounds and experience, teacher backgrounds and experience, student profiles, etc.
- The fourth component on broader institutional factors involves confidential interviews and surveys of district policy makers, key senior staff, union leaders, site participants (teachers, parents and administrators), and external community leaders. These have been supplemented by a careful review of official policy and implementation documents produced by the Board of Education, the Association, the administration and the Design Team, as well as press reports, newsletters, and other public documents.

The chapters which follow present a detailed examination of Pay for Performance in Denver, a summary of findings, and a comprehensive analysis of the impact of the pilot to date. Recommendations are presented in the final chapter.

CHAPTER II

Pay for Performance

A. Introduction

In September 1999, the Denver Public Schools (DPS) and the Denver Classroom Teachers Association (DCTA) charted a new course. They joined to co-sponsor the Pay for Performance Pilot. This pilot is designed to establish a linkage between teacher compensation and student achievement.

The two parties are collaborating to help the district become more focused on increasing student achievement and improving overall professional performance. By the close of the pilot, DPS and DCTA also intend to develop a new salary structure that will be based, in part, upon student academic achievement.

Through the pilot, the parties are committed to a greater emphasis on results in the district. The Board of Education established the Pay for Performance Pilot for Teachers as one of the district's highest priorities. It is also one of the highest priorities of the Association. The parties further demonstrated their joint commitment to results by agreeing to commission an independent, third-party research study on the impact of the pilot.

This pilot and the study are being undertaken during a period of intense national focus on accountability as the means to improve public education at the local, state and federal levels. It comes at a time when an abundance of solutions have been proposed to address problems in American public education. Many of these solutions are being adopted or imposed without sufficient understanding of either the challenges of implementation or the impact—both intended and unintended—on students, teachers and schools.

In this context, Denver's Pay for Performance Pilot may significantly help to shape the discussion and direction of educational policy. The pilot provides a broad and comprehensive experiment with teacher compensation that is linked directly to performance, uses several measures, is supported by labor and management, and takes place in the real-world setting of a large urban school district. Moreover, the results of the pilot are being documented and scrutinized. This is a local pilot with far-reaching implications.

B. Broad Significance

There is a distinct need for educational accountability. However, the rush to implement new programs and initiatives without understanding their functioning or impact—not only in specially selected schools but also in complex school districts —has frequently resulted in failed improvement efforts and significant unintended consequences. The combination of scope and substance provided by the pilot, and the accompanying study of its impact, makes Denver a unique learning laboratory for the nation to examine how Pay for Performance affects students, teachers, and schools, and to explore the reasons for such impact.

The Historical Context

Concerns about educational accountability are longstanding. Particularly since the development of educational testing at the beginning of the 20th century, there have been numerous efforts to measure student proficiency and teacher effectiveness. They have frequently emerged in times of perceived national crises. For example, the launch of the Sputnik in the 1950s generated concerns about America's level of educational preparedness, and the 1983 report, *A Nation at Risk,*[1] drew a link between school performance and the national economic recession.

Pay for Performance, or *Merit Pay,* has an even longer history in the field of education. Researchers Wilms and Chapleau trace the attempt to pay teachers according to their perceived effectiveness back to England in 1710. By 1862, this practice had become a part of the British Revised Educational Code, but by the 1890's linking teacher pay to student achievement had been removed. Educators and inspectors came to believe the practice produced teaching to tests, rote learning, and cheating. One inspector wrote that he observed "children reading flawlessly, but with their books upside down."[2]

Similar efforts were attempted at various junctures in the 20th century. The Nixon administration sponsored a national experiment in the late 1960s, and a series of proposals and experiments were initiated in the mid-1980s in response to *A Nation at Risk.* Called by some, "…as American as apple pie,"[3] each of these experiments generated an initial period of apparent success. This was followed by serious problems and disillusionment, as questions were raised about the narrowing of the curriculum, teaching to the test, and the reliability of the tests and measures being used. Wilms and Chapleau conclude that "Politically driven reforms are nothing more than reflections of public frustration. And, rather than helping to solve root causes of failure, they paralyze us and deflect public attention from reforming educational systems at their core."[4] Or, as Odden has observed, "The '80s saw significant experimentation but few pay systems of merit survive."[5]

The Contemporary Context

In recent years, the country has focused once again on accountability in education. Much of the emphasis has been on developing and implementing educational standards and related assessments. While few states had developed statewide standards in 1990, forty-nine of the fifty states have now generated educational standards. Also, the President of the United States is currently promoting a plan to link federal educational funding to school success in helping students to meet the standards.

As part of the accountability movement, considerable attention is being paid to issues of teacher effectiveness. Teacher training programs are being modified, a host of professional development programs for teachers have been developed, and initiatives to link teacher compensation to teacher effectiveness are emerging. The business community, through the project, "Investing in Teaching," sponsored by the Business Roundtable, the National Association of Manufacturers, and the U.S. Chamber of Commerce—National Alliance of Business, has indicated strong support for a range of experiments around teacher compensation, including the pilot taking place in Denver.[6] Further, while teachers' unions at the local and national level have historically opposed direct linkages between compensation and student performance, some, including some districts represented in the Teachers' Union Reform Network, have promoted union sponsorship of such initiatives.[7] In addition, many local unions, as in Denver, are working with boards of education and district managers to develop customized approaches to promote and reward teacher effectiveness.

Within the range of proposals and experiments concerning teacher accountability and performance, there are numerous differences in approach. In many of the experiments, the focus is on demonstrating teacher skill and knowledge. For example, a Cincinnati, Ohio plan focuses on professional development and creates five career categories for teachers.[8] Some experiments focus on groups of teachers rather than the individual teacher, and explicitly reject the notion that the compensation of individual teachers can be linked fairly to the achievement of their students. Other experiments, such as the Denver pilot, are based on the belief that student achievement is the bottom line of education—and that pay for performance needs to link directly to student achievement.

Despite their variety and differences, the experiments undertaken to date have yet to demonstrate sustained success in improving student achievement in the complex context of an actual school district. Where individual schools and classes have demonstrated improvement, these gains have most often come under highly controlled conditions.

Few comprehensive studies of the results of these experiments have been undertaken. Despite the lack of research regarding actual impact and consequences, more than twenty school systems—from Boston to Dallas—have moved to incorporate variations of pay for performance into district operations. This pattern is also reflected in state and federal policy initiatives.

In response, the National Education Association (NEA) and the American Federation of Teachers (AFT) have stated opposition to linking individual teacher pay to student results. For example, Tom Mooney, President of the Cincinnati Federation of Teachers and an AFT Vice President has said, "I think it is unsound, unprofessional, unethical and no one has shown me a system that's even halfway credible."[9]

With both powerful proponents and opponents, pay for performance and the varying offshoots of the concept are becoming core elements of the American educational landscape. This overall effort is summarized by Adam Urbanski, President of Rochester Teachers Union, who has observed that: "Merit pay is not going to go away until there are some alternatives to it."[10]

C. Colorado and the Denver Area

Colorado has been a locus for attention on performance initiatives. In 2000, former Vice President Gore endorsed "…the Denver Experiment."[11] Douglas County, Colorado has been implementing a form of PFP based on teacher acquisition of skills and knowledge since the early 1990s. Both the Denver Board of Education and the Association have been sensitive to the increasing calls for greater accountability and changes in the public schools.

Some of this is reflected legislatively in the State of Colorado's recently enacted Colorado Revised Statute, CRS Section 22-7-607.5 (formerly known as SB 00-186, Section 19) "Concerning Education Reform, And Making an Appropriation Therefore." Passed on May 28, 2000, it directed the Colorado Department of Education to: (1) develop comprehensive data collection and reporting systems, (2) prepare a school report card for each school, and (3) assist in performance decisions at all levels of school administration. This act gave birth to the Colorado Student Assessment Program (CSAP) and the school report cards which are now important features of the educational agenda in Colorado. Further, it created the "Excellent School Award Program," required additional reading, writing and math assessments, and required each school district to have district-wide and local school accountability committees. The requirement for accountability committees follows other state level charges, such as Collaborative Decision-Making, which date back to 1991.

The above activities, other state level initiatives, such as the Colorado Education Association Regional Bargaining Council's late 1993 Joint Bargaining Council Project on the "Single Salary Schedule and Alternative Compensation," and public concerns regarding accountability and student performance exist in the context of prevailing teacher salary structures. Similar to many districts throughout the nation, Denver compensates teachers through a structure that includes negotiated cost of living increases, compensation based on years of service and considerations based on course credits and/or degrees obtained. All of these factors served as a backdrop for the evolution of Denver's agreement on PFP.

D. Pay for Performance in Denver

Origins

There are as many perspectives on the origins of pay for performance in Denver as there are groups involved in educational reform. Regardless, the Board of Education's discussions regarding PFP culminated at an August 1998 retreat. As reported in a background paper,[12] the retreat produced, "…one cornerstone of the Board's vision…change the way teachers are paid." The intent was "…to link teacher compensation to student achievement." The Board further said that for the plan to be successful, "it must …be fair, competitive and attractive to employees." According to this background paper, prospective criteria for a PFP plan included the following: eliminate automatic raises, link all raises to the achievement of specified goals, have the new pay system create a healthy school environment that puts the focus on student achievement without teachers feeling as though they are competing against each other, measure achievement in terms of student growth or the value added by teachers, improve the competitiveness of teachers' salaries in the Denver metropolitan area; and, finally, be capable of being implemented and financed over the long term.

Pay for performance soon became a significant item in the contract negotiations between the Board of Education and the Association. After an extended period of negotiations, which included the involvement of a third party mediator, the Board and the Association agreed in September 1999 to implement the pilot, identified in the contract as Pay for Performance.

Initial Terms

The terms of Pay for Performance comprise Appendix E of the Bargaining Agreement.[13] Key features, summarized below, include:

- Instituting a two-year pilot.
- Establishing a Design Team composed of two teachers selected by the Association President and two administrators selected by the Superintendent. All four members were released from other duties.
- Charging the Design Team with guiding, overseeing and implementing the pilot. This included authorizing the Design Team to seek outside assistance and to evaluate the pilot.
- Defining the terms for participation. The initial goal was to have 12 elementary schools and 3 middle schools volunteer to participate. The participant schools each required an 85% vote of support of the faculty.
- Building the pilot around teacher-set objectives. Each participant would collaborate with the principal/supervisor to establish two performance objectives.
- Establishing the financial terms for the pilot. These included:
  - The existing salary schedule would be maintained during the pilot.
  - In Year I, the 1999-2000 school year, each participant would receive a stipend of $500 for participation. Also, participants would receive an additional $500 if they attained their first objective and $500 if they attained their second objective.
  - In Year II, the 2000-2001 school year, participants would receive $750 for each objective met.
- Defining the approaches that participating schools would use to measure progress in meeting the objectives.
  - In Approach One, participants would develop and test objectives based on the Iowa Test of Basic Skills, a national norm-referenced test.
  - In Approach Two, participants would develop and test objectives based on teacher developed criterion-referenced tests, or other teacher developed measures.
  - In Approach Three, participants would develop and test objectives based on increases in teacher knowledge and skill.
- Indicating reporting dates for the Design Team. The most pivotal requirement is that the Design Team would issue a report by June 1,

2001 (near the end of the second year of the pilot) which would include a description of pilot effects and would provide recommendations for implementation in the next year. The Board of Education and the Association would subsequently determine whether to proceed with further implementation of Pay for Performance.

In late October 1999, the Design Team announced the first 12 schools to be included in the pilot, based on school-by-school votes taken among the teaching staffs. They were Centennial, Colfax, Columbian, Cory, Edison, Ellis, Fairview, Mitchell, Oakland, Smith, Southmoor and Traylor Academy. All of these are elementary schools. None of the middle schools generated the 85% vote threshold at this time. The Horace Middle School became a pilot school in the 2000-2001 school year.

Revised Terms

Denver's goal was to conduct a pilot characterized by quality and integrity. Accordingly, the sponsoring parties understood that mid-course corrections might be required as part of the pilot's implementation. The major institutional adjustments are delineated in Chapter VIII of this report. The following are select revisions which affected the core construct and intent of the pilot.

By December 1999, it had become apparent to the Design Team through the external assistance provider (CTAC) that modifications were necessary. First, a baseline period for measuring progress needed to be established. Second, more than two years were needed to accurately identify student achievement trends. Third, the final reporting date in June 2001 precluded the use of that year's ITBS scores in analyzing results—due to a lack of data availability.

Beginning in January 2000, the Design Team presented these concerns to the Board of Education, the Association and external supporters. They formed the basis of considerable discussion and collaboration. Several revisions—at policy and operational levels—resulted. These included:

- Extending the pilot to a period of four years.
- Establishing new reporting dates and products. An interim report would be presented in November 2001. A final report would be presented in November 2003.
- Defining the baseline year for study purposes. This would be the 1999-2000 school year.
- Changing the threshold for faculty votes to participate. Participation in the pilot initially required an 85% vote of the faculty. This number was changed to 67%.
- Establishing the need for a group of control schools for study purposes.

In June 2000, additional challenges were identified. These included (1) the need for a succinct, clear statement of purpose for the pilot, and (2) the need to identify a vehicle for directly addressing the development of a new salary structure. These concerns were raised in various forms and settings, including formal and informal discussions between the Board of Education and the Association, written correspondence between internal and external parties, a board retreat on June 26, 2000, and numerous other discussions.

The Board of Education and the Association collaborated again. They agreed to amend the prior agreement. They established a statement of purpose and created the Joint Task Force on Teacher Salary. Both the statement of purpose and the description of the task force were made appendices to the contract.

The formal Statement of Purpose follows:[14]

> *The mission of the Denver Public Schools (DPS) is to graduate students who are literate and who possess the thinking skills and personal characteristics needed for a successful transition to the post-high school experience. Our teachers offer the key link to ensuring that each child reaches their fullest potential. The value placed on the teaching corps is reflected in the financial commitment the district has made to teacher salaries, which is the single largest item in the budget. To establish a structure of salary advancement that recognizes the efforts of teachers in a child's academic success, the Board of Education and the Denver Classroom Teachers Association (DCTA) have initiated a Pay for Performance Pilot. The Pilot has been designed to identify an appropriate method of measuring a teacher's effectiveness in the classroom.*

> *The Pay for Performance Pilot is a learning endeavor in which DPS and DCTA will jointly develop a compensation system based in part on student achievement. To do so, DPS and DCTA have established a Design Team to oversee the pilot and to develop a method for teachers and principals to set academic achievement objectives. The DPS and DCTA will establish a joint task force to design and recommend the salary structure that will support this system.*
>
> *In the fall of 2003, the Design Team will draw together the results of the pilot and the work of the joint task force. The pilot will be evaluated by a third party, the Community Training and Assistance Center, and the results of the pilot will be presented to the Board of Education and the members of the Association.*

In a separate Memorandum of Understanding, the Joint Task Force on Teacher Salary was established. This task force, comprised of representatives from the Association, administration and the community at large, is to function apart from the collective bargaining process. The basic charge for the task force is to "develop and analyze model salary systems for appropriate teacher pay for performance in the Denver Public Schools."[15]

With these revisions, the Denver Board of Education and the Association have taken their initial concept of pay for performance and made corrections in an effort to strengthen the implementation and study of the pilot.

E. Summary

Of all of the current experiments in teacher accountability, it is clear that paying individual teachers based, in part, on the performance of their students, is one of the most controversial and contested. This is largely due to the failed attempts of the past and the skepticism that individual teacher performance can be tied legitimately to student achievement. The concept enjoys both strong support and strong resistance. At the same time, such a model gets at the very core of accountability, as supported by business and community leaders.

The Denver Pay for Performance Pilot is nationally distinct. Key components are described above and referenced throughout this report. Most particularly, the pilot attempts to forge a direct link between teacher compensation and student achievement. In this effort, the pilot cuts across lines of ideology and interest (labor/management, liberal/conservative, Democrat/Republican) through joint board and union sponsorship. By so doing, the pilot moves beyond rhetoric to focus on results.

CHAPTER III

Research Design

A. Overview

Purpose

The primary purpose of this study is to examine the impact of the Pay for Performance Pilot. The study focuses on changes in student achievement between pilot and control schools, across pilot approaches, and with teacher objectives. It examines teacher objective-setting in detail. It also examines the impact of the pilot on school and district-wide practices, as perceived by teachers, site administrators and parents, and considers institutional factors that have affected implementation. The findings from this study are presented in Chapters IV through VIII, followed by a discussion of critical questions and issues in Chapter IX and recommendations in Chapter X. There are four primary areas of study.

Student Achievement

The foundation of the pilot is student achievement. What changes or growth in student achievement have occurred at pilot schools? To what extent do these differ from changes that have taken place at control schools? How can student achievement be compared to teacher objectives? How do other student, school, or teacher factors, or pilot approaches, correlate with student achievement and affect analysis of pilot results?

Objectives

Classroom objectives set by teachers and approved by principals are the cornerstone of Denver's pilot. How do they compare to quality standards? How have objectives changed since the pilot began? To what extent does the quality of the objectives or teachers' success in meeting their objectives correlate with classroom measures of student achievement?

School, Teacher and Student Factors

There are significant differences between the student populations at schools and in classrooms, and among teacher factors such as teacher experience. These factors vary considerably among pilot schools as well as between pilot and control schools. To what extent do these factors correlate with student success? To what extent do these factors affect the various approaches adopted by each school?

Institutional Factors

The pilot is being implemented in a large and complex urban school district, with all of the challenges, internal and external problems and pressures that exist in that setting. What institutional factors have significantly influenced the implementation of the pilot, and how have these factors affected pilot results seen to date?

The study's approach to examining these issues is outlined below. Each chapter of analysis fully describes the methodologies employed.

The Role of the Community Training and Assistance Center

The Community Training and Assistance Center is filling a dual role with regard to Pay for Performance. First, it is providing technical assistance to help assure pilot quality and integrity. Second, it is studying the impact of the pilot. As such, CTAC is in the unique position of a participant observer, developing a case study of Denver's implementation of PFP. There are both pluses and minuses to this approach.

The role of technical assistance provider presents the potential for introducing bias into the examination of study impact, particularly in areas where CTAC has made specific recommendations. This potential cannot be eliminated; rather, it is acknowledged as a possible influencing factor in arriving at the conclusions and recommendations presented here. In conducting the study and presenting the results, however, we have taken the steps most often identified as appropriate for this form of research (Merriam, 1998; Stake, 1991; Yin 1984).[1] First, the study has drawn on multiple sources and has clearly identified those sources in describing what has taken place and in drawing conclusions. Second, CTAC has publicly described its relationship to pilot participants, all of whom have known of our roles. Finally, we have identified the sources of the conclusions, in that the route to these conclusions is clear. All of the quantitative data are a matter of record, and these form the basis of many of the conclusions. The identification of issues and subsequent recommendations are matters of interpretation, and should be seen in the light of CTAC's extensive, albeit outside, involvement in the pilot.

On the positive side, however, this particular kind of study required this level and form of involvement. The Board of Education, Denver Classroom Teachers Association, Design Team and the pilot's funders wanted to know not just what has happened, but also *why* and what needs to be considered next. Therefore, CTAC's active involvement in the pilot has been essential. CTAC played no role in the initial design or structure of the pilot, nor in the original negotiations, but it has been a close observer of subsequent activity. Context is critically important both in determining what has happened and what needs to be done next. This is a strength and a requirement of this study.

B. Student Achievement

Selection of Assessments

The central questions with regard to student achievement are how achievement has changed at the pilot schools, how achievement at pilot schools differs from control schools, and what impact other pilot factors, such as pilot approach or quality of objectives, have had on achievement.

Early in the pilot, the Design Team and DPS departments of Assessment and Testing, and Curriculum and Instruction, created an assessment matrix outlining the appropriate assessments already in use within the district as they pertained to the three different approaches. In a September 2000 report, the Design Team identified 13 assessments that could be used in the different elementary grades, including ITBS, parts of the Colorado Student Assessment Program (CSAP), the 6+1 Trait Writing Sample (Six-Trait), and measures for younger children encompassed within the Colorado Basic Literacy Act (CBLA). Because all pilot school teachers are involved, however, including classroom teachers, special subject teachers, special education teachers, and specialists such as nurses and psychologists, many different measures have actually been utilized in teachers' objectives. The Design Team identified 116 different assessments used by at least one teacher in its June 2000 report.

It is important to note, with regard to assessment, that because the goal is to measure teacher impact on a classroom or group of children, most measures used in objective setting are predicated on student growth rather than comparisons of achievement across groups of students. The state's CSAP, which was designed for other purposes and which does not provide a mechanism for pre- and post-testing of an individual child, was less appropriate for objective setting. According to the Colorado Department of Education, it will soon be possible to examine reading scores from year to year through vertical scaling. Currently, CSAP is used primarily for grade level comparisons and not individual student differences.

For the purposes of our analysis, which is a broad comparison of student achievement across classrooms and schools, most of the 116 assessments noted above are not appropriate. To make broader comparisons, CTAC has selected those measures that appear to be most widely used. Thus, it has focused on the following three assessments: ITBS, CSAP, and Six-Trait Writing.

Description of Assessments

The Iowa Test of Basic Skills (ITBS)

The ITBS, developed by the Riverside Publishing Company, is a norm-referenced achievement battery composed of tests in several subject areas. In the development process, as described by Riverside, all the tests were administered under uniform conditions to a representative sample of students from the nation's public and private schools at each grade level. This process produced the test's battery scores, scales and norms. Different grades were required to take different subtests from year to year. This prevented the comparison of some grades and subtests from one year to the next.

It should also be noted that some students are excluded from taking the ITBS at the principal's discretion. Since this discretion may be exercised differently at different schools, the numbers of test takers and the results may vary. Also, in setting their objectives, teachers are not held responsible for students who do not meet certain criteria; for example, they may have entered a teacher's classroom mid-year, or have been chronically absent. Since these factors do not appear in the database, they are not considered in our analysis. CTAC has conducted analyses on third through fifth grade classes at pilot and control schools. Only one third of the schools were Approach One schools (focused on ITBS), and in any given school there are many specials and specialists who may have used different measures to set objectives. These factors all account for some differences between school results reported in this study and individual teachers' success in meeting their objectives.

There are several reasons for including ITBS in the study. First, Approach One specifies the ITBS for teacher objective-setting. Second, at the beginning of the pilot it was the most widely used assessment in the district. Third, it is norm-referenced and developmentally scaled such that student growth can be measured from one year to the next.

To assess the impact of the pilot, CTAC calculated the change in normal curve equivalent (NCE) scores between the Spring of 2000 and the Spring of 2001 for each student on each test (reading, language, and math). This allowed the study to use the paired comparison of means methodology to test whether the mean change in NCE scores is zero. A change of zero means that a child has performed as expected, or grown an average amount, for one year of instruction and development. A positive change in NCE means that the child is now performing at a higher level than expected given the previous year's score. A negative change in NCE means that the child is performing at a lower level than expected. A basic assumption of the paired comparison of means test is that the observations are independent, an assumption which is usually violated when subjects are grouped within schools and classrooms.

In addition to the paired comparison of means, hierarchical linear modeling (HLM) was used to compare the results of each approach to the control group. The HLM model uses a realistic assumption—that there is correlation between student scores within the same school and classroom.

HLM techniques are described in further detail in Chapter VI, as part of the discussion of CSAP analysis. Chapter V identifies issues that impact the ITBS analysis, examines the demographics of the samples, and presents the results.

Colorado Student Assessment Program (CSAP)

CSAP was developed for the State of Colorado by CTB/McGraw-Hill. CSAP tests are based on the Colorado Model Content Standards and are used for accountability purposes across the state. The Colorado Model Content Standards represent the fundamental knowledge and skills that the state of Colorado expects students to possess at various intervals as they move through their educational careers. According to the Colorado Department of Education, CSAP tests consist of a mix of constructed response (25%) and multiple choice items (75%). Item response theory methods were used for test analyses, scaling, equating, to form the item selection process, and to place both multiple choice items and constructed response items on the same scale. Scale score cut-points were set that define four performance levels—Unsatisfactory, Partially Proficient, Proficient, and Advanced. The standard setting was conducted using the Bookmark Standard Setting Procedure.

CSAP has grown in importance during the past few years because it is used by the Colorado Department of Education to rank schools. School-by-school results are published in the newspapers and announced on television. Colorado is like most states in that it does not, as yet, adjust its statistics for differences among children and classrooms, with the result that schools where the children primarily come from low-income families, are members of minority groups, or are non-native speakers of English, are consistently ranked lower than other schools. As the state's largest urban school system, Denver consequently has a disproportionate share of schools rated as "low" and "unsatisfactory." For these reasons, CTAC has included CSAP in its analysis.

Administration of CSAP in Denver follows state guidelines, which have not been consistent over the past two years. The grade levels at which the tests are given, and the content of those tests, have fluctuated from year to year and content has not been completely standardized. Nor can CSAP subtests yet be used for measuring individual student growth from one year to the next. For example, the two subtests of CSAP that the study focuses on in the analysis—reading at the third and fourth grade levels—still cannot be used to show student growth. That is, the nature and scoring of the tests prevents using the third grade test as a starting place for measuring the growth of fourth graders. The Colorado Department of Education is working towards vertically scaling CSAP tests.

CSAP does present one testing advantage over ITBS, which is that it is based on state standards. To the extent that the curriculum and instruction at a given grade level address the state standards, CSAP should provide a closer link to teacher performance than a more general test.

The differences between CSAP and ITBS have led to significantly different analyses. ITBS allowed for measuring student growth from one year to the next, which eliminates the need to test for many child-level characteristics. Using multiple test results from the same child provides a strong control, as most child and family characteristics change little from one test administration to the next. Lacking that form of accounting for child-level differences, the analysis of CSAP has focused on accounting for child level factors through statistical methods, and comparing these to classroom level factors, to test the extent to which these factors correlate with assessment outcomes.

Two different sets of analyses were conducted. First, regression analysis was used to compare children in pilot classrooms to children in control classrooms in order to examine differences between groups. The test scores of pilot school students in years 1999-2000 and 2000-2001 were compared to scores of control school students for the same years, including reading, writing, math (2000-2001 only) scale scores, and the reading and writing standards scale scores by grade where possible.

Second, the study's analysis of CSAP included using hierarchical linear modeling, the multilevel analysis method described above. HLM is used because it allows the study to partition the variation in children's scores into between-child variance (to be explained by child-level characteristics) and between-class variance (to be explained by classroom-level characteristics) and ensures an accurate estimation of the effects. This analysis fits into one of the most common multilevel analysis models, which allowed the study to examine child outcomes as a function of both child and classroom/teacher factors.

6+1 Trait Writing (Six-Trait)

The 6+1 Trait Writing system was developed at the Northwest Regional Education Laboratory (NWREL). Six-Trait is used district-wide to improve and assess students' writing skills. The Six-Trait analytic rubric used in classroom instruction is also used to score the assessment. This assessment is required for all students in the fall semester and administered again in many schools in the spring. The six writing traits are: Word Choice, Sentence Fluency, Conventions, Ideas, Organization, and Voice. Each of these traits is scored independently by teachers on a five-point scale.

For the purpose of this study, students' Six-Trait scores from the Spring 2000 and Spring 2001 administrations were analyzed.

The Six-Trait Writing assessment produces rubric scores for individual students based on the six writing traits. This scoring system is considerably simpler than the scores for ITBS and CSAP, but does allow for some quantitative analyses. A variety of analyses were conducted using a change score—the difference between their scores in Spring 2001 and Spring 2000. In order to assess the degree of change across a distribution ranging from –4 to +4, independent sample chi square tests were used to compare various groups at each grade level.

Colorado Basic Literacy Act (CBLA)

Through this act, the district employs a range of assessments at the earlier grades. Many of these are used for the youngest students, where the results of an individual administration of the assessment are less likely to be representative of actual student knowledge. The study reviewed the results of these assessments. However, the number tested, differences in how the tests are used, differences between and among the different tests, and the fact that they, like Six-Trait, are teacher-scored, all suggested that these tests were less useful for the broad comparisons being presented in this report. Like Six-Trait, many of these assessments are useful instructional and assessment tools in the classroom, but Six-Trait provides a somewhat more consistent and useful measure for comparative purposes across schools.

C. Objectives

In the pilot schools, each teacher writes two objectives. These are approved by the principal and form the basis for evaluating classroom results. Objective writing is seen as a central component of the pilot.

Several approaches were used to evaluate the quality of teacher objectives. The original plan for the study was to create a series of classifications into which objectives could be grouped, including difficulty, quality, subject matter, whether objectives were met or not met, pilot approach, and relationship to school, teacher, and student factors. However, several methodological and practical considerations arose, causing a redesign of the original classification scheme. First, more than 80% of teachers met their objectives in both of the pilot's first two years, reducing variability to the point where few valid comparisons could be made statistically. Second, it became clear both that too many objectives did not fall clearly within a particular category, and that such categorization would produce many sub-groups too small to analyze.

Quality of Teacher Objectives: The study used several sets of data to evaluate overall objective quality: (1) rubric scores for each teacher's two objectives over two years, 1999-2001; (2) the summary of met/not met objectives over two years, 1999-2001; (3) a comparison of objectives to the school plans in 2000-2001; (4) a comparison of pilot school objectives to control school objectives, 2000-2001; and (5) achievement data on the ITBS and CSAP administered to all pilot schools for two years.

Rubric Evaluation of Objectives: To gauge the rigor and overall quality of the objectives, a four-point rubric was developed based on the traits of learning content, completeness, cohesion, and expectations. The traits for quality educational objectives were derived from a review of teacher planning guides found in the ERIC system, the district scope and sequence, which contains subject standards for grades K through twelve, and the elements listed on the form provided by the Design Team to the teachers. Four levels of performance were established as a way to rate individual objectives. The levels of performance are

- Level 4–Excellent
- Level 3–Acceptable
- Level 2–Needs Improvement
- Level 1–Too Little to Evaluate

All objectives were read and scored. Six hundred eighty-four (684) objectives in year one and 784 objectives in year two were reviewed. These analyses are described in detail in Chapter IV.

D. School, Teacher & Student Factors

Several data sets were used to determine school, teacher and student factors and are described briefly below. These include:

- Surveys
- Interviews
- Teacher demographics and experience from teacher databases
- Student demographics and behavior from student databases
- School factors from student, teacher and school databases

Surveys

Data Collection

CTAC conducted surveys in each of the pilot's first two years, including teachers and administrators at the pilot schools (1999-2000 and 2000-2001), a random sample of teachers and administrators at the control schools (2000-2001) and a random sample of parents at the pilot and control schools (2000-2001).

In the first year, surveys were distributed to all teachers and administrators at the pilot schools. As new pilot schools were added, this initial set of surveys was also distributed at the new schools. These surveys collected early perceptions of PFP, both to parallel baseline data on student achievement and to provide a baseline for gauging changes in perception. A total of 420 surveys were distributed to the 12 elementary schools and the one middle school pilot. Returned surveys represented a response rate of 83%.

In the second year, a revised survey was conducted at these initial pilot schools. In an effort to identify changes in perception, this second survey asked teachers to respond to many questions similar to those asked the first year. It also specifically asked whether they perceived differences in school activity in the second year as a result of the pilot. The response rate for the second year pilot survey was 91%.

In addition, modified questionnaires were sent to a sample of 660 control school teachers and administrators (response rate 37%), and to a sample of 1,200 pilot and control school parents (response rate 12%).

Surveys were anonymous, but teachers were asked to identify their schools, level or type of teaching (such as classroom teacher, special subject teacher, special education teacher, etc.), and such factors as their length of service in the district and at their current school. Surveys asked questions in the following general areas: pilot goals, pilot support, teacher objectives, implementation, professional development, and impact.

Analysis

In analyzing the data, CTAC tested all variables for significance using ANOVA and chi square analyses. These analyses are presented in Chapter VII, and focus on comparisons across the two years of the pilot, the three pilot approaches and teacher factors such as longevity.

Individual and Group Interviews

Data Collection

In addition to the surveys described above, CTAC conducted more than 100 interviews during the pilot's first year, and more than 200 interviews in the second year. The range of interview subjects included teachers, administrators and parents at the school sites, as well as central administrators, board members, association leaders, and outside observers. These interviews serve to explain and elaborate upon the results of the surveys, as well as to suggest responses to many critical questions as to context, history, and perception.

As an example, the following is a breakdown of the 213 second year interviews by position.

Role in District	2000-2001 Number Interviewed
Association Leaders	10
Board Members	9
Central Administration	12
Design Team Members	5
External Community Members	11
Other Site Staff	4
Parents:	
Pilot schools (32)	
Control schools (11)	43
Principals:	
Pilot schools (14)	
Control schools (10)	24
Teachers:	
Pilot schools (74)	
Control Schools (21)	95
Total	213

Teacher Demographics and Experience

Data Collection

DPS maintains a teacher database indicating a variety of factors, including race/ethnicity, educational background, and number of years experience. These factors were correlated with student achievement results and school results according to teacher identification number. Such a correlation has not been conducted before in Denver, as indeed it has not in most school systems, so the task of linking the data was far from straightforward. Significant issues arose in associating students with their teachers using unique teacher identification numbers, as discussed in Chapter VIII.

Analysis

The study included such categories as length of years in school and district, degrees held, rubric scores on their objectives, and whether they met or did not meet their objective(s), as factors in its analyses. For example, the study explored whether the length of a teacher's experience might show a greater correlation to achievement than whether the student was in a pilot or control school. The results of these comparisons to student achievement are contained within the Chapters V and VI. Similarly, the relationship between teacher factors and their objectives was also explored. This relationship is discussed in Chapter IV.

Student Demographics and Behavior

Data Collection

Student demographic variables are included in the analysis of student achievement, and were obtained from the student database. Aggregate student factors are considered school factors, and are taken into account when considering differences among schools. Attendance figures were only available by school rather than by student, so these were only considered as school factors. Since some students miss a considerable amount of school time, this could be a significant missing variable.

Additional data were provided similar to the data shown in the school report cards published annually. These factors include school size, population, attendance rates, percent of students on free/reduced lunch, percent of students who speak Spanish as their first language, percent of students who speak another language other than English as their first language, and enrollment.

Analysis

The main purpose for determining school factors is to correct or account for them in the statistical analyses. Student results may be analyzed according to participation in free or reduced lunch programs, for example. This factor is then taken into account when comparing schools. The results of these analyses, and the methodologies employed, are discussed in Chapters V and VI.

E. Institutional Factors

Data Collection

Data and information about institutional factors were derived from several sources.

Documentary Data

Documentary data were collected from the following sources:

- Design Team: The Design Team made its files generally available, providing primary source data on many aspects of pilot inception and implementation. This included the Design Team's own semi-annual reports, correspondence and meeting minutes, training outlines and materials, correspondence, and other documents.
- Administration, Board of Education, and the Association: Documents requested from DPS included Board News and press releases, descriptive material on particular aspects of the pilot, and internal newsletters and communications. DPS also maintains considerable information concerning PFP and other topics on its website.
- Local and National Press: Press coverage and editorials, both on local activity and more broadly on other attempts at merit pay and pay for performance were obtained from a variety of sources, including Education Week, ERIC, Kappan, the Business Roundtable, and others.

Interviews and Surveys

As described above, these primary sources provide considerable insight as to the actions of different entities and how they are perceived by different people in different parts of the system. Interviews, in particular, were used to explore perceptions of purpose and impact, to gauge the understanding and involvement of different departments and individuals, and to contrast differing viewpoints at different points of time.

CTAC Involvement

The third primary source of data concerning institutional factors comes from discussions of various issues with board members, officials from both DPS and DCTA, Design Team members, members of the corporate and philanthropic communities, and teachers, principals and parents at the sites.

Analysis

A project or pilot can only be successful if it can be implemented. Experience has repeatedly demonstrated the widely varied impact different implementation strategies and approaches in different institutional settings have had on results, even when the programs or strategies being implemented were similar. Accordingly, the study has paid close attention to issues of implementation—both how things are done and how they might be done to be more successful. The sources described above combine to provide the study with a rich and varied range of information as to institutional issues and their impact on the pilot.

The sources above are used in concert, so that conclusions regarding the perception of institutional factors and the impact these factors have had on the pilot are drawn from several sources. These factors are discussed primarily in Chapter VII, where the study reports on the perceptions from the schools, and in Chapter VIII, where the study reviews major institutional decisions and conditions.

CHAPTER IV

Objectives: Linchpin of the Pilot

A. Introduction

> *At the center of the Pay for Performance Pilot is the method teachers and principals will use to set student achievement objectives. This method must be data driven, credible, and fair if it is to be implemented successfully by all teachers and all principals upon completion of the pilot. If the objective system is proven to work, Denver Public Schools and Denver Classroom Teachers Association will decide to use it to grant teachers their annual pay increases.*[1]

The objectives are the centerpiece of the pilot. Two yearlong instructional objectives, developed individually by teachers, form the basis of the extra compensation system being piloted by Denver schools. Teachers in the pilot schools write two objectives, select a measure, establish a growth target, and indicate their expectations of attainment (in numbers or percentages of students). If the teacher and his or her principal determine that the objectives have been met, extra compensation is awarded.

In the educational setting, instructional objectives relate what the teacher will teach and what students will learn. In educational literature, teacher lesson plans, and teaching materials, one finds yearlong objectives, unit objectives, and lesson objectives, each describing the instructional intentions of the teacher. There is not one prescribed method or research-based model for writing objectives to be found in education literature (except for a short period in the seventies when behavioral objectives were in mode). However, there is research, such as that done by Dr. Madeline Hunter,[2] to indicate that well-constructed lessons, including objectives, contribute to student learning.

A review of models suggested for use by teachers[3] indicates that instructional planning includes: (1) what will be taught (concepts, skills, etc.) and (2) how students will demonstrate learning (assessments, products, etc.), as well as (3) teaching strategies. Teacher-written objectives usually reference the agreed upon curriculum for the grade level and subject. These instructional agreements may be found in state, district, or school standards and benchmarks; scope and sequences or curriculum guides; school planning documents; and/or course outlines. In a well-aligned instructional system, teachers will have access to and make use of appropriate and approved assessments for the agreed upon curriculum, including ones that benchmark student growth and/or provide diagnostic help, as well as annual standardized assessments.

Clearly, instructional objectives already carry a heavy load in the educational setting. Embarking on a course that uses teacher-written objectives as measures of teacher performance and compensation adds a significant level of complexity to a teaching practice already rich in preference. However, as the Design Team's project plan denotes and as their diligent work with teachers demonstrates, objectives hold a key position in the Pay for Performance Pilot—the nexus between what students learn and what teachers earn. Two years into the pilot, in the Spring of 2001, two school officials expressed what many of those involved with the pilot have come to realize: writing appropriate, measurable objectives for student achievement *and* compensation is difficult.

> "When we entered into this, I didn't see the difficulty in a fairly simplistic objective-setting process. I can't get over that objectives are so hard to write."

> "I'm more aware of the complexity of the effort to tie—and validate the tie—between setting objectives and performance pay."

Teacher leaders also view the teacher objectives as a significant element of the pilot—the connector between the district's need to raise student achievement and the local teacher's need to maintain a level of autonomy in his or her own classroom. However, as the project has progressed, teacher leaders are realizing the potential for reforming the system around teaching and learning:

> "Now I see PFP as a major systemic reform. Before it was just a pay thing, and then it became a teaching and learning thing, but if we can align the system around teaching and learning…"

> "[PFP] will have an impact on the teaching and learning process: focus, data-driven [objectives], alignment, staff development for teachers, and the idea of doing an assessment at the beginning of the year."

Yet teachers, teacher leaders, and principals have reservations about the fairness of teacher-written objectives and their potential to improve teaching and learning:

> "It is nice to get the money, but I was already doing what PFP does. The goals have not changed a bit—but how they are measured might have changed. Compensation incentive has not had an impact on student achievement."

> "At the beginning there was some confusion over expectations. The goal setting process should be more standardized. There is a danger that teachers will set easy goals if they know they are being assessed on meeting them."

> "…the project, as it is now with individual objective writing, could pit teachers against one another: those with high-performing students versus those with low-performing students or high school teachers versus elementary teachers."

This sample of interview responses demonstrates the range of thinking about objectives and PFP—from the realization that PFP has the potential to create systemic reform to the observation that PFP could create dissension or allow some teachers to take an easy route to the notion that while PFP means extra compensation, it will not lead to a fundamental change in practice or impact student achievement.

As the linchpin of the pilot, that which connects student achievement and teacher performance, teacher objectives have been the focus of a training program developed and delivered by the Design Team and others, as well as the subject of scrutiny for this research study. Early on, in designing the study, questions arose about how best to evaluate the quality of an objective. While education literature and practice contain "how to" advice and models for teachers, as well as taxonomies of objectives for the test maker, there is not a research base to say that one objective is superior to another in terms of getting results. Secondly, because additional compensation programs based on teacher-developed objectives constitute a mostly untried approach, there is little to assist either the staff developer or the evaluator in the quest for a perfect objective.

In the first year of the pilot, the Design Team members who would assist teachers and principals in writing successful objectives faced a short timeline, both in recruiting a representative number of pilot schools and in assisting teachers to write measurable objectives that could result in their obtaining the additional compensation. The district's teachers and principals had an objective-setting process already in use as part of the teacher appraisal system; thus, this style of objective-writing prevalent in the organization seems to have influenced the first efforts of objective-writing for PFP. The logic of this choice is apparent—teachers were familiar with the teacher appraisal system and using it reduced potential concerns about the unknown by building on the known. However, the Design Team also emphasized the importance of baseline or starting data, both for making good decisions about objectives and in establishing reasonable expectations.

In the second year of the pilot, the Design Team refined this approach based on learnings from the first year, emphasizing greater precision in assessment and higher expectations for students. A form was developed that acted as a heuristic for teachers in thinking about and developing objectives. Teachers interviewed in Spring 2001 noted that "There was more clarity in the goal writing this year over last," and that "[the Design Team] and the school are much better with writing objectives."

Other interview and survey data show that teachers appreciated improvements in the training between years one and two, but that they still want greater clarity, more assistance in writing "reasonable objectives," and more staff development related to student learning. Teachers and principals both still have many concerns about PFP that they would like to see addressed (see Chapter VII).

B. Quality of Teacher Objectives: Methodology and Results

In evaluating the quality of teacher objectives, the study uses several data sets: (1) rubric scores for each teacher's two objectives for 1999-2000 and 2000-2001; (2) a summary of met/not met objectives for 1999-2000 and 2000-2001; (3) a comparison of objectives to the school improvement plans in 2000-2001; (4) a comparison of pilot school objectives to control school objectives for 2000-2001; and (5) achievement data on the ITBS and CSAP administered to all pilot schools for two years.[4] The fundamental questions about teacher-written objectives are as follows:

- What are the traits of a quality objective and how are they best described?
- Is there a relationship between the quality of the objective written by the teacher and the achievement of students?
- Is there a relationship between whether a teacher meets his or her objectives by the measures and parameters he or she has set and student growth on an independent, standardized measure, such as the ITBS, that measures general growth?

FIG. 4-1

Traits of Quality Educational Objectives

Trait 1: Learning Content
Content is that which the teacher will teach and the student will learn. Quality learning content is significant to the subject or discipline, appropriate to the student level, and rigorous in thought and application. Content choices should reference agreed upon standards for the subject and grade level.

Trait 2: Completeness
A complete expression of an educational objective includes: the student population to be taught; the objective with learning content; the assessment; the rationale for selecting the objective; any baseline data available to show prior knowledge and/or skills; the expected attainment; and finally, the evidence that persuades the teacher that the objective has or has not been met.

Trait 3: Cohesion
Cohesion refers to the logic and unity among the elements and demonstrates that rigorous thought and careful planning have taken place in the development of the objective. It gives a sense of the whole over the parts.

Trait 4: Expectations
The complete learning objective demonstrates that the teacher understands both the student population and individuals to be addressed and holds high expectations for each student as well as for himself or herself.

FIG. 4-2

Performance Levels

Level 4 – Excellent: The teacher objective meets all of the criteria.

Level 3 – Acceptable: The teacher objective meets basic criteria with some lack of completeness and/or cohesion.

Level 2 – Needs Improvement: The teacher objective meets some of the criteria, but is incomplete and/or lacks cohesive thought.

Level 1 – Too Little to Evaluate: The teacher objective does not meet the criteria; may show a lack of understanding or effort.

- Is there a relationship between teacher objectives and school improvement plan goals and objectives?
- Did the objective-writing effort in the pilot schools differ substantially from that of teachers in control schools?

While some patterns are emerging from the data, these questions will continue to serve as a basis for study as the district enters the next phase of the pilot.

Rubric Evaluation

As a method of evaluating the rigor and overall quality of the objectives, a four-point rubric was developed by a CTAC educator panel based on the traits of learning content, completeness, cohesion, and expectations. (See Figure 4-1)

The traits for quality educational objectives were derived from a detailed review of teacher planning guides found in the ERIC system, the district's scope and sequence, which contains subject standards for K through twelve, and the elements listed on the heuristic provided by the Design Team to the teachers. Next, four levels of performance in meeting the criteria (four traits) were established. (See Figure 4-2)

Combining the criteria and levels of performance, the rubric in Figure 4-3 provides the tool for describing teacher-written objectives for the purpose of the study.

Applying the rubric, a panel of educator-evaluators read and scored second year and then first year pilot school objectives, resolving any discrepancies with a second review and discussion. The chart in Figure 4-4 demonstrates that the majority of objectives for years one and two fall into Level Two of the rubric. Level Four scores increased in the second year as Level Two scores declined. The discussions in Chapters V and VI will show that there is a positive correlation between the quality of the teacher's objective in 2000-2001, as measured by the rubric, and student growth on district measures.

Still another way to view the rubric data is to compare changes of rubric scores between years one and two for teachers who have two years of scores. Figure 4-5 shows a summary of the

FIG. 4-3

Rubric for Describing Teacher Objectives

Level of Performance	Descriptors for Performance Levels
4: **Excellent**	The teacher states clearly what the students will learn, expressing completely and coherently all elements of the objective, including the assessment, and demonstrating high expectations for students. There is a strong sense of the whole.
3: **Acceptable**	The teacher refers (i.e., from a skill section in a book or test or a program acronym) to what the student will learn but may lack thoroughness in addressing the elements of the objective or in making clear the relationship or unity among the elements. The student expectations may seem somewhat conditional or low.
2: **Needs Improvement**	The teacher has attempted to address most of the elements of the objective but may not have stated what students will learn, (stating the objective as an assessment goal rather than a learning goal). Expectations for student growth are low.
1: **Too Little to Evaluate**	The teacher does not address the objective in a manner that shows either an understanding of the task at hand or an effort to complete the task as requested. Objectives may place too many conditions or exclude too many students from the proposed outcome to be reliably assessed.

increases, decreases, and no changes in the rubric scores of teachers from the first to second year. Approximately three-quarters of the teachers present in the Denver system (and whose objectives were available for both first and second years) either increased in rubric score or made no change. Finally, of those writing objectives for both years, 24% to 27% decreased in score.

Most teacher objectives were scored lower on the rubric than was originally anticipated. This outcome is attributable largely to one key trait of the rubric—learning content—that is missing from many teachers' objectives. Where learning content is included in an objective, it is, for the most part, generally referenced, such as reading or mathematics. More content may be implied in the name of the assessment or in the instructional strategies the teacher has identified, if he or she has done so. The expectation trait of the rubric also proved to be a secondary pitfall, particularly in the first year, but understandably so since teachers were searching for a "reasonable" growth target to set in the new system.

While there are exceptions—for example, there are Level 3 and 4 scores among these objectives—most objectives are "partial," stating how student learning will be measured but omitting what students will learn, except by implication. These partial objectives are referred to in the analysis as "assessment-focused objectives" as opposed to "content-focused objectives," which the rubric describes and which are critical for gauging what students are learning.

In considering why content, or what students will learn, is critical in the written objectives, it is helpful to look at a current model for instructional planning. In *Understanding by Design,*[5] Wiggins and McTighe describe a "planning backwards" model with three steps: "(1) identify desired results; (2) determine acceptable evidence; (3) plan learning experience and instruction." Desired results are not test results, but rather "what students should know, understand, and be able to do." It is interesting to note what the authors say about identifying desired results:

> "What should students know, understand, and be able to do? What is worthy of understanding? What enduring understandings are desired?
>
> "In this first stage, we consider our goals, examine established content standards (national, state, and district) and review curriculum expectations. Given that there is more content than can reasonably be addressed, we are obliged to make choices."

From this premise, the authors go on to describe a method to filter learning content in order to establish priorities for what students will learn— "by design" and not happenstance. Much thinking about content occurs before the second step where assessment or evidence gathering is undertaken. These writers also consider the planning of assessment to be very important (to be undertaken before planning learning activities), but not without reference to what is to be learned.

This is not to say that classroom teachers in Denver did not or do not have content in mind when they are writing their objectives. They surely do. However, since the content is not written into the objectives, what is being taught—or assessed—is not clear, nor is the connection between what is being taught and learned and pay for performance. Adding a content piece will not be complicated—it could be as simple as a reference to the district curriculum standard—and it will add valuable information as the pilot

FIG. 4-4

Tally of Two Teacher Objectives by Rubric Level for All Pilot Schools, 1999-2001

Rubric Level	99-00 Number*	99-00 Percent**	00-01 Number*	00-01 Percent**
4	6	1%	70	9%
3	165	24%	177	23%
2	418	61%	425	54%
1	51	7%	104	13%
NA	44	6%	8	1%
	n = 684		n = 784	

* numbers may differ because of data missing or staff changes
** percent of the total number of objectives evaluated
n total number of objectives available to evaluate
NA not available: blank or incomplete objective forms

proceeds in determining if additional compensation can lead to improved learning.

The rubric-based evaluation has provided a new set of questions about the proper model of objectives that are intended to serve a dual purpose: (1) identification of what is to be taught and learned, the purpose of most instructional objectives written by teachers, and (2) provision of the basis of additional compensation, a new and different purpose.

Emerging questions include:

- Is it possible to write a dual-purpose objective that is clear about what is to be learned and still measurable for compensation?
- Will the connection between teacher performance and student learning become lost with the content unexpressed or indirectly expressed?
- Will the assessments in use become the teaching content in lieu of clear direction about the content to be addressed?
- Are teachers currently writing two sets of objectives, one set to guide instruction, another for compensation?

These questions will need to be addressed as the pilot continues into its next phase.

Analysis of Assessment-Focused Objectives and Content-Focused Objectives

As explained previously, a review of the teacher-written objectives in the Denver PFP project shows that most teachers, in the first two years of the pilot, wrote objectives proposing student growth on a test measure as the instructional objective, which we have referred to as "partial" or for this discussion, "assessment-focused." The following samples from the 2000-2001 pilot schools show some typical assessment-focused objectives:

- 80% of identified students will show one year or more growth in reading as measured by pre- and post-testing on Developmental Reading Assessment (DRA)/Qualitative Reading Inventory (QRI) assessments.
- 98.5% will demonstrate growth on the following math assessments of the Brigance: 4A, 7A, and 10A.
- 3 of 23 students will move from the first quartile to the second quartile (on ITBS Reading).
- 80% of non-waived students, who took the Fall pre-test and who are in attendance for 85% of the school year, will demonstrate five to seven months growth on the Spring ITBS Reading.

As these samples indicate, assessment-focused objectives are written to show the teacher's intent to improve student performance on a test over the year and to indicate what assessment will be used to show improvement. In many cases but not all, objectives set an expectation level for achievement (in numbers or percentages of students who will improve on the test). What will be taught and learned is not the focus of an assessment objective though content may be implied broadly as the content of the assessment, such as reading or math.

Still other teacher objectives from the 2000-2001 pilot schools are established on content or skill subsections of an assessment like Six-Trait Writing and thereby give more information about the learning content. For example:

- Of the children scoring 2, 3, or 4 on the Organization section of the fall Six-Trait Writing sample, 80% will improve by at least one level according to the Spring sample of the test.
- 70% of students in grades 2 and 3 will increase their decoding of high frequency words by 15 words.

FIGURE 4-5

Number and Percent of Rubric Score Changes from 1999-2000 to 2000-2001 for Teachers with Two Years of Objectives

Objective One				Objective Two			
School	# / % Rubric Increase	# / % Rubric Decrease	# / % Rubric No Change	# Teachers with Two Years Objectives	# / % Rubric Increase	# / % Rubric Decrease	# / % Rubric No Change
All Pilot Schools	90/38%	56/24%	91/38%	237	86/36%	64/27%	87/37%

- 75% of students enrolled by 9/30/00 will increase one proficiency level on the data analysis and problem-solving portion of the DPS Math Assessment Spring 2001.

By referencing subsections of assessments, such as organizational skills for writing, decoding of high frequency words, and problem solving and data analysis, teachers have provided more information about what students will learn.

Some teachers did write objectives with emphasis on content. For example, teachers of special classes or electives, such as art, music or physical education were often content-specific in their objectives:

- Based on pre- and post-tests, 75% of the afternoon early childhood education students, in attendance 85% of the school year, will recognize numerals 1-10 and recognize the basic shapes of circle, square, triangle, and diamond.
- 70% of students will recognize by sight or sound violin, trumpet, clarinet, flute, snare drum, bass drum, harp, and piano. They will be able to write a paragraph about the instruments.
- 80% of identified students will increase their level of understanding and application of color theory and color mixing as demonstrated on a teacher made spring assessment (sponge painting color wheel).
- I will increase my competence in teaching kids to determine importance in reading passages. This will be evidenced by the improvement of two thirds of my class in at least one skill level on the Exemplar rubric.

In these objectives, content is the focus of the objective and is stated clearly enough that an outsider reading the objective can, for the most part, form a mental picture of what students are learning and how the teacher will know that they have learned.

Why Assessment-Focused Objectives?

In considering why assessment-focused objectives prevail in the PFP project rather than content-focused objectives, which are available in district curriculum documents and which teachers must use to guide their lesson planning, the following organizational factors seem to be influential:

- Since the final assessment determines the additional compensation, it is eminently practical to write the objective to the assessment as well as to narrow the expectation of attainment by eliminating many of the conditions over which teachers have little control (attendance, etc.).
- Teacher performance appraisal (between principal and teacher) documents show objectives written in the assessment-focused mode, as evidenced in the control school documents where these appraisals are in use.
- In survey and interview data, numerous teachers in the pilot schools report having been told that they could join the pilot and get additional compensation for doing "what they were already doing."
- The district scope and sequence core document does not have assessments aligned with the standards in such a way that teachers can readily make use of them for PFP. Secondly, it is not clear that all Denver teachers were able to access adequate information on items or subsections of the ITBS or CSAP in order to know what specific content and skills are assessed.
- Materials provided to teachers emphasized consistency and clarity in measurement, important to the success of the compensation program, more than standards and curriculum objectives.

Why Not Assessment-Focused Objectives?

Since a core intent of the pilot is to improve student learning, the question for the project is how best to establish a connection between pay for performance and the actual teacher practices that will improve learning. At this point, the objective is that nexus. So an objective that, at a minimum, communicates the key learning(s) as well as the assessment is essential. With the prevalent assessment-focused objectives as they now stand:

- One can read several hundred objectives and not come away with a strong sense of what is being taught and learned in the participating schools. The content must be inferred in a general way (i.e., reading, math) from the assessment.

- It is not evident, since there are few references to the agreed-upon curriculum, that the objectives are grade-level appropriate or of high priority to the school and district.
- The relationship of teacher objectives to school improvement plan objectives or district goals is minimal.
- One could infer that the content to be taught is that tested on the standardized assessment.
- Opportunities for thoughtful teacher discussion and reflection on alignment and focus—factors likely to improve student achievement—may be lost.
- Assessment-focused objectives, while representing a potentially useful element of system-level targets—state, district, or school, do not seem as helpful to classroom teachers.

Examples of Content-Focused Objectives

To show how reading objectives may also be more content-focused, the following yearlong objectives were developed based on fifth grade Reading Standards 5, 6, and 7 found in the McRel K through twelve Content Standards.[6] The 75% attainment is chosen for demonstration purposes and assumes, at least, a year's growth. Though the following examples show the final measure as the ITBS, other and/or additional measures are desirable.

- 75% of students will improve in their use of the general skills and strategies of the reading process (preview, purpose, predicting, context clues, phonetic and structural analysis clues, and speed) as measured by reading passages on the Iowa Test of Basic Skills (from grade four administration to grade five administration).
- 75% of students will improve in their use of reading skills and strategies to understand and interpret a variety of literary texts (literary forms and genres; plot; character; point of view; inference; and theme) as measured by reading passages on the Iowa Test of Basic Skills (from grade four administration to grade five administration).
- 75% of students will improve in their use of reading skills and strategies to understand and interpret a variety of informational texts (defining characteristics of a variety of texts, organizers, parts of book, viewpoint, main idea, supporting details, and organizational patterns) as measured by reading passages on the Iowa Test of Basic Skills (from Grade 4 administration to Grade 5 administration).

These concepts and skills form the basis of a year's worth of work in fourth grade reading. In the Pay for Performance Pilot, teachers may choose to focus in on fewer skills for the purpose of their own evaluation. With an analysis of student scores from the previous year or from a classroom diagnostic test given at the beginning of the year, a teacher can determine where instructional emphases throughout the year will shore up weaker areas while continuing to build on strengths that students have already demonstrated. For example, a teacher may find that most students are reading narratives with understanding, but need more work on making sense of informational texts. That teacher may then write a PFP objective that focuses on the main idea, supporting details, and organizational patterns in informational texts.

The early analysis of what students know also provides teachers with a preliminary identification of which students will need differentiated instruction. Finally, for the supervising principal or parent reviewing the objectives, it is clearer what students will be learning throughout the year.

Making the Objective Count for Instructional Purposes and PFP Purposes

The pilot links student learning to teacher compensation, and since the objective is the nexus between these two, it seems logical that the development of objectives for the year needs attention. The findings of the study and the work of the Design Team show that it is not necessarily easy to write instructional objectives that also meet compensation goals. In surveys and interviews, teachers have asked for more help—both in constructing objectives and in working with data. The following will assist teachers and principals:

- Time provided for teachers at the end and beginning of the school year to review the agreed upon curriculum, establish priorities, analyze the available assessments, and any and all information available about how students have performed to date—minimally, test information broken out into clusters, and preferably, item analyses, student learning records, or other diagnostic information.

- A framework for teachers to use in developing quality instructional objectives that is based on agreed-upon traits like those seen in Figure 4-1.
- The district standards/scope and sequence available as the content source book for classroom objectives.
- Expected growth targets or levels of attainment established so that teachers have measurement standards with which to work. The Design Team currently has a study group addressing this issue.
- Priorities for staff development and classroom resources established based on the objectives set and the identified needs of students.

Over the long term, the district must join in with the schools in a more formal alignment process to help teachers with appropriate objectives, assessments, student expectations, and teaching resource—thereby providing students with a learning support system that is clearly understood by everyone, including parents. But for now, the annual setting of objectives for PFP within a school can be a powerful opportunity for dialogue about improving student achievement.

Teachers Meeting or Not Meeting Objectives

Most teachers in both years were able to meet the objectives they set for themselves and receive the additional compensation. Out of 341 teachers in 1999-2000, 325 met the first objective and 303 met the second objective; in 2000-2001, 387 out of 423 met the first objective and 382 met the second objective.[7]

Since most teachers met their objectives, there is not enough difference to compare met/not met data with rubric scores though there were some interesting findings about rubric scores, met/unmet data, and student achievement that are discussed later in this chapter and in Chapters V and VI.

Comparison of Pilot School Objectives with Control School Objectives 2000-2001

Teacher objectives from twelve of the control schools (year 2000-2001) were reviewed for comparison to the pilot schools. The teachers in the control schools wrote their objectives for the teacher appraisal system. On a one-page form, the teacher is asked to write two objectives that address district and/or school goals. A third (or more) objective(s) may address other areas. Most teachers write a professional development objective for their third objective. They are then asked to assign a weight to each of their objectives. Some do it evenly—33.3% per objective; others weigh each objective differently—40%, 35%, 25%. Another section of the form asks for the strategies that will be used to meet objectives. Finally, there is a section for the appraiser to rate the performance of the teacher on a four-point scale, from "outstanding" to "unsatisfactory."

Most of the objectives are written in the assessment-focused style observed in the pilot school objectives. The percentage of attainment designated by the teacher in the control schools is generally high—over 75% mostly and occasionally set at 100%. Examples of typical objectives follow:

- 100% of the students who are with me for the entire literacy block will show one year's growth as measured by the QRI II in the Spring of 2001.
- 95% of the students who attended all year and achieved a Partially Proficient or Proficient score on the third grade CSAP reading test, will make at least one year's growth and score Partially Proficient or Proficient on the fourth grade CSAP.

As with the pilot schools, some teachers in the control schools do write more content-focused objectives, but generally, they center their objectives on improving student performance on an assessment. There is little use of baseline data evident on the forms and expected growth is mostly a year's growth on the assessment. Expectations in terms of percentages appear to be higher than pilot schools, but there are occasionally exclusions or disclaimers, such as "students who attended all year."

FIG. 4-6

Percentage of Teachers Meeting Objectives 1 and 2 in Years 1999-2000 and 2000-2001

School	Total # Teachers 1999-2000	# Meeting Objective 1 1999-2000	# Meeting Objective 2 1999-2000	Total # Teachers 2000-2001	# Meeting Objective 1 2000-2001**	# Meeting Objective 2 2000-2001**
Centennial	35	33	31	39	38	37
Colfax	25	24	24	28	26	27
Columbian	23	19	16	25	25	22
Cory	27	27	27	26	24	24
Edison	29	28	28	34	31	29
Ellis	35	35	35	35	32	32
Fairview	27	27	27	33	32	31
H. Mann*				50	40	43
Mitchell	33	33	33	35	31	34
Oakland	35	34	32	37	31	31
Smith	34	27	13**	33	32	28
Southmoor	10	10	9	17	16	15
Traylor	28	28	28	31	29	29
All Pilots	341	325/95%	303/88%	423	387/91%	382/90%

* Entered Pilot in 2000-2001
** Data not complete

There are thematic similarities within schools that show either grade-level planning or school planning. One principal of a control school did not send copies of objectives, but reported in a memo that all teachers in the school had the same two objectives, increased proficiency on CSAP reading and math.

Influence of School Plans on Teacher-Written Objectives

A review of the school improvement plans of the pilot schools showed them to be uniform in format with sections to address special populations and with the same or very similar goals. The goals in the school improvement plans are often more content-specific than the teacher objectives. However, in comparing the teacher objectives to the school improvement plans, it is difficult to see a direct match, possibly because many objectives are nonspecific on content. The exceptions tended to be in the areas of Six-Trait Writing and number sense, both of which are explicit in most plans and show up specifically in many teacher objectives. In one case, a staff development element (writing) of the plan was very influential in teachers' objectives. Figure 4-7 shows a sample of five of the schools, labeled here as Schools A, B, C, D, and E.

When asked about the influence of the school improvement plan on PFP, interviewed staff usually responded that the school plan set a focus for the year, and thus, in that manner, it influenced their choices for PFP. Most interviewees did not see a direct connection between the plan and the objectives (although in one school all teachers agreed to and set the same objective), but more of an indirect influence.

The school improvement plan objectives are similar by school, with only limited customizing by the schools. There are apparently longstanding

FIG. 4-7

Sample Comparison of 2000-2001 Teacher Objectives to 2000-2001 School Improvement Plan Objectives

School	General Characteristics of Teacher Objectives
A	Out of 76 objectives (38 x 2), most (31) addressed reading as an improvement on a reading test; the second largest group of objectives (12) addressed writing as an improvement on a rubric; and the third largest group (11) addressed mathematics as an improvement on a math assessment. Other objectives addressed language arts skills (10), subject specific skills (3), personal/social skills (5), and teacher staff development (2). There is a tendency for teachers to set objectives in two different areas (i.e., reading and math), but several teachers wrote both objectives in the reading area. The most straightforward connections to the school plan are those objectives referencing the Six-Trait Writing process. The objectives appear to be modeled more on the PFP template and examples than on the school plan, and measurements are those on the district-approved list supplied during the PFP workshop.
B	Of 66 (33 x 2) objectives, more than half (37) address reading; 12 objectives address mathematics; five address subject or special skills; three each address language arts and personal growth skills; two address writing; and one was incomplete. In general, the learning objectives are expressed as growth on an assessment. The assessment is either named or a generic "pre-post-test" term is used. Several math objectives address number sense, which is indicated as a goal in the school plan. However, the objectives appear to be selected from or modeled on the PFP examples provided for teachers rather than the school plan.
C	Of 68 objectives (34 x 2), 24 address mathematics; 21 address reading; nine address subject or special skills; eight address writing; and three each address language arts and personal skills.The objectives are expressed as growth on an assessment, modeled on the examples provided for PFP. The objectives show a high level of expectation with few conditions attached to them. There seems to be an emphasis on math in this school, including number sense, computation, data analysis, and problem solving.
D	Of 100 objectives (50 x 2), 35 address writing; 19 each address reading and subject skills; 15 address personal skills; six address mathematics; two address language arts; and one addresses staff development. Three objectives were not clear or incomplete. A beginning- of-school workshop on writing apparently was a strong influence on teacher choices in this school. The content of the objectives range from general improvement on the Six-Trait Writing test to a degree of improvement on subskills such as word choice. Also several of the reading objectives address vocabulary, which was also part of the workshop. Since this is a middle school, it is logical that there would be more subject-specific objectives. Though more learning content can be derived from this school's objectives, they still emphasize growth on an assessment.
E	Out of 66 objectives (33 x 2), most addressed reading using a uniform objective that states: "80% of non-waived students who took the fall-pretest and who are in attendance for 85% of the school year, will demonstrate 5-7 months growth on Spring ITBS Reading." This objective appears to have been required or agreed upon for the first objective for most teachers; the second set of objectives also includes a widely used uniform objective for math similar to the one for reading. Twenty-three objectives address mathematics. One objective each addressed language arts, subject skills, and personal skills.

General Characteristics of School Improvement Plan	Degree of Match*
The plan makes use of a district template and apparent suggested objectives for the categories of reading, writing, and mathematics. For each of these three categories there are specific strategies for staff development, teaching approaches, gifted students, multiple opportunities, and library. Generally, the content of the reading, writing, and mathematics instructional approaches in the plan are more specific on content than the teacher objectives in each area (i.e., weekly numbers sense instruction, a variety of literary genres, Six-Trait Writing), but without measurements. The lack of alignment between strategies and measurements probably means that the school plan was not particularly useful to teachers for writing PFP objectives.	Partial match in the area of Six-Trait Writing.
The school plan makes use of a district template with some difference in the math staff development strategies. There are no measurements suggested for the school plan content, though improved measurement is a strategy in the plan. The lack of alignment between strategies and measurement probably means that the school plan was not helpful to teachers for writing PFP objectives. The math section specifies regular work in number sense, which is echoed in some of the mathematics objectives of teachers.	Partial match in the area of number sense.
The school plan makes use of a district template with some differences, notably the omission of number sense, which is included in other school plans. However, number sense is the content of several teacher objectives. There are no measurements for the plan strategies.	No match evident.
The plan, which is modeled on a district template, specifies that there will be schoolwide training in Step-Up-to-Writing. The actual strategies, however, identify the use of technology in writing and the use of guided practice in writing in all classes. The most evident connection between the school plan and the teacher objectives is the choice of many teachers, teaching a variety of disciplines, to set objectives for writing and vocabulary (an element of the writing workshop). This is the influence of the writing staff development component of the plan rather than the writing strategies section of the plan.	Partial match from influence of writing staff development.
The school plan is based on a district template. There is no evidence of influence on the teacher objectives except for the general categories of reading and math.	No match evident.

*Degrees of Match are explained below:

Comprehensive	Clear evidence in specificity of objectives, language of the objectives, and priority of objectives that the school plan was referenced by a substantial number of the teachers.
General	The objectives selected by a substantial number of teachers reference the school plan categories as priorities (i.e., reading, ELA) or list them as a rationale for their choices.
Partial	Specific parts of or strategies from the school plan are referenced by some teachers.

concerns with the school improvement plans. A central administrator indicates, "The school improvement plan was more an act of compliance. It has been historically. There is not a lot of alignment at the sites. It's not as widespread as it could be." Also, the evaluation section of the school improvement plans is not completed and the timeline says "ongoing." It may be difficult to equate a PFP system with very specific assessments and a nine-month or less timeline with the school improvement planning system which is ongoing with no published assessment. However, this is an issue of alignment that the district will want to address.

Perceptions about the Quality, Credibility, and Influence of Objectives on Student Achievement

The interviews with teachers, principals, and other staff, as well as the survey data and comments, provide a window into the perceptions about objectives as an element of the PFP pilot. While there were a range of responses and feelings, all were thoughtful, and even in cases where the person did not support the concept of pay for performance, he or she was helpfully analytical about issues. From these data, three perceptual themes emerged that are worth considering here.

Teachers are doing what they have always done. While this is the prevalent attitude of the interviewed teachers toward the objective setting process, it is also apparent in many of the principal interviews. If any ground is yielded here in terms of change, it is in the areas of focus and measurement, as in the claims to be "more focused," "focused earlier in the year," and "measuring growth more carefully or more frequently." According to these interview data, formal pre-tests or baseline tests were not part of the practice until PFP, so it seems that setting objectives with pre-test information alone would lead to the conclusion that this process is different; however, teachers often have access to other student data—cum folders, running records, informal inventories, and conversations with previous teachers—and feel that these are helpful in determining student needs. Survey data support the statements of interviewed teachers:

- 69% of teachers disagree/strongly disagree with the statement "I am doing things differently as a result of PFP." and 79% agree/strongly agree with the statement "I am doing what I have always done to meet the objectives of PFP."
- 61% of classroom teachers and 75% of administrators agree/strongly agree with the statement "There is a close relationship between objectives under PFP and teacher evaluation."
- 48% of classroom teachers and 63% of administrators agree/strongly agree with the statement "PFP has led to greater focus on student achievement at my school."
- 86% of classroom teachers and 86% of administrators agree/strongly agree, "Most teachers are using student achievement data to develop objectives."
- There is a close response among teachers about whether student achievement has stayed the same or increased (44% to 47%). In addition, 72% of administrators believe it has increased.
- The relationship of school practices to student achievement, the relationship of teacher practices to student achievement, the relationship of professional development to student achievement, the competition and cooperation among teachers have remained the same according to survey percentages.

Teachers do not see PFP as a major change in classroom practice as indicated by both surveys and interviews. The general look of the control school and pilot school objectives supports this observation as well. Since improving student achievement by rewarding improved teacher performance appears to have been a major focus of the project, "business as usual," if true, may be a counterforce in improving student achievement in the district.

Interviewees speak of increased focus, earlier focus, focus on certain students as a result of PFP, and just less than half of the teachers indicate agreement with the statement that PFP has led to greater focus on student achievement on the survey. There could be some different uses of the word "focus" occurring (i. e., personal planning as opposed to student achievement) but it seems not.

More, earlier, or different focus on learning by teachers (especially within schools or grade levels) is a well-recognized element of improved student achievement. So gaining focus is both doing something different and doing something that matters.

The objective-setting process is unfair because

- Of the diversity of students in the school, in the classroom, in the district;
- Some teachers set lower expectations and/or manipulate the outcome;
- Teachers are judged on a single measure;
- Teacher performance for pay is based on measuring student performance;
- Standardized tests are a poor measure of student performance.

Again, the issues of fairness come up from teachers and principals predominantly, but a few central administrators are concerned about the lack of calibrated measurements. The central staff were also concerned that the objectives teachers set may not relate to CSAP standards.

There is concern about fairness also among the parents interviewed, legitimate concerns about minority students becoming increasingly taught by less able teachers as the best teachers start to move to higher performing schools. Secondly, there is concern among parents that minority teachers will be adversely affected. The survey data contain the following responses on fairness:

- 69% of classroom teachers and 88% of administrators agree/strongly agree that "fair objectives can be set by all teachers."
- 56% of classroom teachers and 50% of administrators agree or strongly agree that "principals are ensuring that teachers set equally challenging objectives."
- 56% of classroom teachers and 100% of administrators agree/strongly agree that "it is possible for all teachers to meet their objectives."
- 81% of classroom teachers and 88% of administrators believe that "most teachers will meet their objectives this year.
- 75% of teachers and 88% of teachers agree or strongly agree that "variations in classroom composition" make it difficult to set fair objectives.

There is agreement between the interview data and the survey data on the issue of diversity of students and fairness.

Most of the fairness issues identified are real ones with potential solutions; others are probably based on mistrust of the administration or other teachers, or on a concern about one's capacity to work with students who need differentiation. For example, the lack of consistent, calibrated measurement; the lack of foresight in planning for the special teachers; and the lack of outside supervision for teacher test administration have potentially short-term solutions. Developing trust for one another will come over time as the pilot progresses and problems are solved. However, finding ways to teach all children effectively will require staff development.

Among all the interviews, only one person spoke of "closing the learning gap between minorities and whites" as a desired outcome of PFP and as an indicator of the success of the PFP. While several school officials noted the excessive "excuse making" about difficult-to-teach students, the school staff questions "their accountability for students who are achieving below standard" at the beginning of the year. This concern underlies many of the fairness issues that come up among school staff, and it shows up in the way teachers have designed the objectives.

There are many ways to resolve the measurement issues, and a committee is currently working on these measurement/gain issues. However, there are also myriad classroom strategies for differentiation in successful use around the country. Teaching all students, intervening effectively with those who fall behind are subjects of professional development, and such staff development should be part of any plan to adjust or calibrate the measurements.

There needs to be more training and professional development for teachers and principals. There is a bit a of a disconnect between "I'm doing what I've always done" and "I need more help." However, teachers and principals, in general, though conceding that the introductory sessions by the Design Team were helpful, want more. There seems to be a consensus that the Design Team was better organized and more precise in the second year orientations. Yet teachers and principals indicate that more professional development is needed.

Principals speak of lack of follow through, of the project just being dropped in their laps as the arbiter of the quality of the objectives, and of increased workload. Some teachers speak of not hearing back from their principals about their objectives and of a lack of follow-through. With a few teachers there is some sense that they should be doing something different in the classroom, but they are not hearing what it is.

The teachers who are responding the most positively to the PFP objectives have or have had in the past some helpful staff development (as in the writing workshop or the literacy training). Also, teachers and principals mention falling back on one another for assistance, either formally, with grade level teams, or informally. Though some teachers feel that PFP has interrupted the team culture of their schools because teachers are measured separately and the outcome is confidential, most do not see this as a problem. Principals indicate a need for assistance with providing effective classroom observations and feedback. In general, there is recognition from the Board of Education to the school practitioners that professional development is currently inadequate and needs to be strengthened.

The word "reasonable" in reference to the objectives was repeated frequently in the interviews. It may mean "attainable" in this context—not too difficult to be achieved but beneficial to students. Survey responses show that teachers and administrators believe that objectives are both attainable and challenging, but they need more training both to set and to meet objectives:

- 78% of classroom teachers and 86% of administrators agree/strongly agree that "most teachers are setting objectives that are both attainable and challenging."
- 78% of classroom teachers and 63% of administrators agree/strongly agree that "teachers need training and support to set objectives."
- 68% of classroom teachers and 75% of administrators agree/strongly agree that "teachers need training and support to meet objectives."

Teachers respond that support in implementing PFP, in improving classroom instruction, and in understanding student achievement data have all remained the same over the two years of the pilot. Clearly, the fact that many staff members would approach an initiative like PFP with a "business as usual" attitude is an indicator that there is not a clear connection between improving classroom practice in order to improve student achievement.

Comparison of Objective Rubric Scores and Met/Not Met Data with Student Achievement Data

Still another way to look at teacher objectives is to compare their quality—rubric scores—with the student achievement data. A thorough analysis of these data and the methodology used for analysis are included in Chapters V and VI. However, there are findings from these analyses that are notable here:

- The Spring 2000 ITBS mean reading NCE increases as the quality of the teacher objective increases—from 41 for the lowest category (rubric score 1) to 52 for the middle two categories (rubric scores 2 and 3) to 71 for the excellent category (rubric score 4). Additionally, the percentage of students performing at/or above grade level increases in a similar manner: 33% for the lowest category; mid-50% for the middle categories; and 84% for the excellent category.
- The Spring 2001 ITBS shows the change from the previous year's administration. For teachers with rubric scores of 1, 2, or 3, the mean change in student reading scores is close to zero (a year's change) and sometimes negative, but for teachers in the highest category (rubric score 4), the increases were 2.6 for Objective One and 1.4 for Objective Two. These numbers indicate that the students of teachers who scored 4 on the rubric achieved more than a year's development, on the average, while the students of those teachers who scored in the lower categories gained a year's growth or lower.
- For excellent objectives (rubric score four), the students achieved whether the objectives were

met or not met—a 2.8 NCE increase if the objectives were met and a 2.4 NCE increase if the objectives were not met. This illustrates the importance of articulated learning content and high expectations. However, for lowest rubric score, unmet objectives are associated with decreases in NCEs; for rubric score 2, unmet objectives are associated with a 2.2 increase while met objectives are associated with a negative 0.4 NCE. Achievement levels for met and unmet middle (rubric scores 2 and 3) quality objectives were mixed.

- On the CSAP, the quality of teacher objectives also proved to be a significant classroom level predictor of fourth grade reading achievement. The analysis indicates a significant positive relationship between the quality of teacher objectives, no matter the approach, and student test scores. A one level increase in the quality of teacher objectives (from rubric score 1 to 4) is associated with a 13 point increase in student reading scale scores on the average.

These early findings support the premise that a high quality teacher objective, one demonstrating all of the rubric traits, leads to greater student achievement. In the succeeding years of the pilot, if all teachers write content-focused, or more meaningful, objectives, more data will become available to determine how the quality of an objective set by the teacher impacts the outcome for students.

C. Summary and the Challenge Ahead

This chapter has discussed the search for a meaningful way to discern the quality of the teacher objective, particularly since the objective is seen as significant to both the success of the pilot and to the improved performance of students. A four-trait, four-level rubric was developed for the purpose of assessing the quality of the objectives, which could then be compared to student outcomes. However, for the organizational reasons discussed in the previous section (expediency, past practices, and lack of aligned curriculum support materials and school plans), most teachers did not reference the content to be taught in their objectives, thereby generating a preponderance of mid-level rubric scores. Also, because most objectives were met, the line of met/unmet data were less helpful than anticipated in distinguishing the quality of the objectives.

The study also explored teacher and principal perceptions about the objectives through survey data and interview data and found that (1) many teachers believe that they are doing what they have always done; (2) there are several fairness issues, unintended, that need to be addressed, especially in the areas of expected growth of students and the appropriate assessment of special teachers and specialists; and (3) teachers and principals need more assistance with the objective setting and more professional development, particularly that which will help teachers with addressing diverse learners and principals with providing classroom feedback to teachers.

Finally, the study found that higher quality teacher objectives (rubric score 4) were associated with higher performing students on the ITBS reading and on the CSAP fourth grade reading. This indicates that fully developed instructional objectives will likely move the district toward the goal of improving student performance. Clearly, there is a challenge and opportunity ahead for the Denver Public Schools to:

- Address the fairness issues identified by teachers, particularly the student growth issue and the special teacher and specialist issue.
- Align assessments and materials using the newly developed curriculum standards and help schools create plans that are implemented and evaluated annually.
- Provide professional development for teachers and principals that focuses on the teaching and learning issues in the district.
- Embrace the diversity of the students by providing the training and resources that will allow all teachers to succeed with students who require more or different instruction.
- Encourage and assist with the development of PFP objectives that reflect the dual nature of the pilot: objectives aligned with the standards for improved student achievement that are also fair and measurable for compensation.

CHAPTER V

Iowa Test of Basic Skills

A. Introduction

Student achievement is central to the Pay for Performance Pilot. Each of the pilot's three approaches to pay for performance includes a component of student achievement. Approach One teachers set objectives that are assessed through the Iowa Test of Basic Skills; Approach Two teachers set objectives that are assessed through a range of teacher-developed or criterion-referenced assessments, including CSAP. Approach Three is based on teacher acquisition of skills and knowledge, but also includes student achievement goals measured in one of the above ways. Special subject teachers, special education teachers, and specialists such as nurses and psychologists, who make up more than half of the "teacher" population in Denver, have written objectives that vary both by approach and the particular school or schools they work in. In all, the Design Team counted more than 116 assessments used by at least one teacher for objective-setting in the pilot's first year.

For the broader question of the pilot's impact on student achievement, it was neither possible nor desirable to attempt to measure impact in every assessment. Rather, based on the district's broadest measures of student achievement, the study examined whether there were differences between schools and teachers according to the following:

- Between pilot and control schools
- Between one pilot approach and another
- According to identifiable student factors
- According to identifiable teacher factors
- According to identifiable school factors

B. Assessments

This chapter focuses on the Iowa Test of Basic Skills (ITBS). This is one of three assessments that were used as part of this study.

Iowa Test of Basic Skills

The ITBS, developed by the Riverside Publishing Company, is a norm-referenced achievement battery composed of tests in several subject areas. In the development process, all of the tests were administered under uniform conditions to a representative sample of students from the nation's public and private schools at each grade level. This process produced the test's battery scores, scales and norms. At the start of the pilot, Denver students in grades K through eight were required to take the ITBS reading test. Students in grades two, five and eight were also administered the ITBS math test.

Form M of the ITBS (1993) is the version currently used within the district. High school students in grades nine through eleven take the Iowa Test of Educational Development (ITED). The Spanish language equivalent of this test used in Denver is the Aprenda. Since ITBS was specified in Approach One, was the most widely used assessment in the district at the start of the pilot, and is developmentally scaled such that student growth can be measured from one year to the next, the ITBS became the anchor assessment for measuring overall pilot impact on student achievement.

Colorado Student Assessment Program (CSAP)

CSAP was developed for the State of Colorado by CTB/McGraw-Hill. CSAP tests are based on the Colorado Model Content Standards and are used for accountability purposes across the state.

CSAP has grown substantially in importance since the pilot began, to the point where it appears to be the most widely watched and discussed assessment in the district and in the state. While CSAP was initially given at different times and in different subjects for students in different grades, and was thus difficult to use for comparative purposes, a more orderly use of this assessment system is developing. Because of its importance in Denver and Colorado, we have compared CSAP results broadly across approaches, pilot and control schools, despite the fact that relatively few teachers set their objectives based on CSAP. A complete discussion of the study's CSAP analysis is provided in Chapter VI.

6+1 Trait Writing (Six-Trait)

The 6+1 Trait Writing system and assessment is a classroom-based instructional program driven by a rubric. Its primary purpose is to assist classroom teachers to improve student writing. The virtues of Six-Trait are that it is a useful instructional tool and that it is widely used throughout the district. The constraints with Six-Trait include uneven administration and, in most cases, classroom teachers score their own students. For instructional purposes this is appropriate, but for the purposes of consistency of scoring and objectivity, it raises questions of validity and reliability. Despite these implementation issues, Six-Trait provides a measure that can be compared across schools.

A general note on methodology. The differences between these assessments in the way they are constructed, used at the classroom level, and implemented across the district suggested different approaches to analysis. When possible, ITBS and CSAP were analyzed using similar statistical methods, but because ITBS allows the measurement of individual student growth, student growth was the focus for the first analysis. ITBS also allows for some additional analyses, including value-added modeling. These analyses will be ongoing during the balance of the pilot, as two years of data provide only a minimum with which to begin to build an appropriate model. While CSAP does not yet measure student growth, it provides scores according to a number of more specific standards. Therefore, the emphasis was placed on student, teacher and school factors, and the extent to which differences in performance can be explained or accounted for. In Six-Trait, which produces less quantitative data for each child, chi square analysis was conducted. Analysis of all of these assessments will be greatly enhanced during the balance of the pilot by the addition of more years of data, as well as by addressing issues of consistency in the district's data.

It is likely that analyses produced in this report will differ in some respects from data produced by the district for several reasons. First, much of the data received for analysis from this most recent school year (2000-2001) was in a fairly raw form. Data often have to be examined and cleaned, and present multiple questions to be addressed at the school level. While the district has continued to do this as a part of its normal functioning, the study has used data from a particular point in time to proceed with its analysis. Second, many decisions have to be made in the course of analysis, such as where to establish data inclusion cut-off points. Do you include classes that have fewer than 10 students, or fewer than 15? If only 50% of students at a grade level have scores for a particular exam, do you compare that with a school where 100% of students have scores. These decisions are made in the normal course of analysis, and not all of CTAC's analysis decisions will be identical to those of the district. For all of the above reasons, readers should expect minor variations in the statistical summaries presented in this report and those presented at varying points in time by the district.

C. ITBS Methodology

Using the ITBS data, we were able to answer a number of questions about the pilot study. We can determine whether the control and pilot groups are performing at grade level relative to a representative national sample. We can also tell whether each group of students, on average, achieved a year's increase in development between Spring 2000 and Spring 2001. We can compare the three pilot approaches to the control group and determine whether the approach groups performed better, worse, or the same as they would have in the absence of the pilot program. Finally, we can measure the correlation between student success and teacher objectives, both the overall quality of those objectives, and whether objectives were met or not met.

Two attributes of the ITBS make it useful in assessing changes in student achievement. The test scores for each grade are normalized against a representative national sample of students, and are also transformed into normal curve equivalents (NCEs) such that they have a normal distribution centered at 50.

To assess the impact of the pilot, we calculated the change in NCE scores between Spring 2000 and Spring 2001 for each student on each test (reading, language, and math). This allowed us to use the paired comparison of means methodology to test whether the mean change in NCE scores is zero. A change of zero means that a child has performed as expected, or grown an average amount, for one year of instruction and development. A positive change in NCE means that the child is now performing at a higher level than expected given the previous year's score. A negative change in NCE means that the child is performing at a lower level than expected. A basic assumption of the paired comparison of means test is that the observations are independent, an assumption which is usually violated when subjects are grouped within schools and classrooms.

In addition to the paired comparison of means, we also used hierarchical linear modeling (HLM) to compare the results of each approach to the control schools. The HLM model uses a realistic assumption that there is correlation between student scores within the naturally occurring hierarchies of classroom and school, (e.g. students exist within the classroom, classrooms exist within the school). This technique also allows us to control for differences between the pilot and control groups that may impact achievement levels.

At the elementary level,12 schools self-selected into three treatments, or approaches. Approximating the overall demographics of the pilot schools, 36 elementary schools were selected to serve as a control group. Elementary school students are assigned to one teacher. However, a student could have more than one teacher if a classroom teacher left during the school year or if a student changed school or classroom. For purposes of the study, all student results were based on the teacher they had in October 2000.

At the middle school level, one school serves as the pilot school and the remaining middle schools as the control schools. Middle school students have multiple teachers, making it more difficult to assign full credit for a given student's achievement on the ITBS to any one teacher. In addition, teacher assignments are available for the pilot school but not for the control schools, so we are limited in the HLM models to controlling for the correlation within pilot schools only.

Because of differences in teacher assignment noted above, the middle school analysis differs from the elementary school analysis, so these are presented separately.

Qualifying Statements

The analysis plan called for each school (both control and pilot) to administer the ITBS reading, language, and math tests to all eligible students beginning in spring of the first grade through fall and spring of the eighth grade. Spring 2000 to Spring 2001 tests are used in this analysis.

Elementary School Data

At the elementary level, the maximum number of subjects produced by this approach is 6,394 Control, 1,077 Approach One, 756 Approach Two, and 794 Approach Three students. According to data available, seven control schools and no pilot schools followed the complete testing schedule. Most of the schools did not administer all of the tests to all of the grades, sometimes testing all students in selected grades plus a handful of students in non-tested grades. The small numbers of students in non-tested grades are likely to be either under assessment for potential learning difficulties or students new to the school, and thus would not be representative of all the students in their grade. Overall, approximately 75% of the students were tested. For this analysis, we chose to exclude grades within a school in which less than 65% of the students were tested. Figure 5-1 lists the schools and the grades tested for reading, language, and math. Excluding the children in non-tested grades results in sample sizes of 4,909 Control, 886 Approach One, 675 Approach Two, and 630 Approach Three students for the reading test. For the language test the sample sizes are 3,755; 886; 248; and 173 respectively, and the math test sample sizes are 3,622; 809; 508; and 339 respectively.

As Figure 5-1 shows, only a few Approach Two and Three schools implemented the language test, and then only in some grades. The math test presents similar problems. Comparisons for language and math are presented, but since we do not know whether there are systematic differences between the schools that did and did not administer tests, these results should be evaluated with caution.

Middle School Data

The pilot middle school did not administer ITBS exams to the eighth grade in Spring of their seventh grade year, limiting the middle school analyses to the sixth and seventh grades. For the reading test, the pilot sample includes 318 students, for the language test 313, and for the math test 192. Sample sizes for the comparison schools are 6,061 reading, 5,913 language, and 5,304 math. Thirteen students have ITBS test results assigning them to the pilot school, but the school does not list them as taking any courses. These students are included in the comparison to the control schools, but are excluded from teacher level analyses.

Exclusions

The data do not contain any authoritative indicator of whether a child's scores should be counted or excluded. Children who do not speak English well enough to understand the exam and children with physical or learning disabilities that would interfere with successful test-taking should be excluded because their test scores are not a valid reflection of their abilities. There are different approaches to this issue. We chose to present an analysis of all students and adjust for language ability, presence of any disability, and presence of an ITBS exclusion code. A second analysis was then done using our best estimate as to who should be excluded. Children who had a test exclusion code on any ITBS (Spring 2000, Fall 2000, or Spring 2001), who had a LAS code of 1 or 2, or who had a LAU code of A or B were dropped from the Non Excluded analysis. Over time, a further analysis of excluded students may be developed jointly with DPS.

The human resources database contained information on teacher degrees but not on licensing and educational background data for all teachers. Further, the match between student and teacher was not complete. Not every teacher number attached to a student appears in the human resources file, and not every student in the sample was attached to a teacher. Some of these discrepancies are due to changes in the teacher force—teachers retiring or moving, for example—

FIG. 5-1

ITBS Testing in Spring 2000 and Spring 2001

Elementary Schools	School	ITBS Reading				ITBS Language				ITBS Math			
Group		Grade											
		2	3	4	5	2	3	4	5	2	3	4	5
Approach One	Colfax			x				x				x	
	Oakland	x	x	x	x	x	x	x	x	x	x		x
	Smith			x	x		x	x	x	x	x	x	x
	Traylor	x	x	x	x		x	x	x		x	x	x
	% Tested	50%	50%	100%	75%	25%	75%	100%	75%	50%	75%	75%	75%
Approach Two	Columbian		x		x		x		x		x		x
	Edison	x	x	x	x						x	x	x
	Fairview	x	x	x	x	x	x	x	x		x		x
	Centennial	x	x							x	x		
	% Tested	75%	100%	50%	75%	25%	50%	25%	50%	25%	100%	25%	75%
Approach Three	Cory	x	x	x	x						x		
	Ellis		x	x	x						x		
	Mitchell		x		x	x	x		x	x	x		x
	Southmoor	x	x	x	x					x	x		
	% Tested	50%	100%	75%	100%	25%	25%	0%	25%	50%	100%	0%	25%
Control Schools	Asbury	x	x	x	x	x	x	x	x	x	x	x	x
	Ashley			x				x				x	
	Bromwell	x	x	x	x		x	x	x		x	x	x
	Cheltenham				x								x
	Doull		x	x							x		
	Ebert	x		x				x				x	
	Follis				x								
	Force	x	x	x	x	x	x	x	x	x		x	x
	Garden Place		x	x	x		x	x	x		x	x	x
	Gilpin				x				x				x
	Godsman			x	x			x	x				x

while others simply represent missing data. These may be considered common data gaps in any system attempting to use administrative information for analytic purposes. Over time, the district's ability to fill these gaps will result in a stronger analysis.

D. Comparison of Pilot and Control Demographic Factors

Elementary Schools

Ideally, the elementary school control group should be representative of the whole district. In

FIG. 5-1 CONTINUED

ITBS Testing in Spring 2000 and Spring 2001

Elementary Schools	School	ITBS Reading				ITBS Language				ITBS Math			
Group		Grade											
		2	3	4	5	2	3	4	5	2	3	4	5
Control Schools Continued	Goldrick			x	x				x				
	Greenlee			x	x			x	x				
	Gust	x	x	x	x	x	x	x	x	x	x	x	x
	Lincoln	x	x	x	x		x	x	x		x	x	x
	McMeen		x	x							x		
	Montclair		x				x						
	Moore	x	x	x	x		x	x	x		x	x	x
	Newlon		x	x	x		x	x	x		x	x	x
	Philips	x		x	x	x		x	x	x		x	x
	Remington			x	x			x	x				x
	Rosedale	x	x	x	x	x	x	x	x	x	x	x	x
	Schmitt	x	x	x	x	x	x	x	x		x		
	Steck	x	x	x	x					x	x	x	x
	Steele	x	x	x	x		x	x	x		x	x	x
	Slavens	x	x	x	x						x		x
	Teller		x		x				x		x		x
	University Park	x	x	x	x	x	x	x	x	x	x	x	x
	Valverde		x	x	x		x	x	x		x	x	x
	Whittier	x	x	x	x	x	x	x	x	x	x	x	x
	Maxwell	x	x	x	x	x	x	x	x	x	x	x	x
	Amesse			x	x			x	x			x	x
	Holm		x	x	x		x		x		x		x
	Kaiser	x	x	x	x	x	x	x	x	x	x	x	x
	Samuels	x	x	x	x		x	x	x		x	x	x
	McGlone												
	% Tested	50%	67%	83%	83%	28%	53%	67%	72%	28%	61%	58%	72%

At Least 65% of Students Tested, by School, Subject, and Grade

order for the lessons learned from the pilot study to be applicable to the district, it would also be desirable that the three approach groups be similar to the control group. However, for most of the student and teacher characteristics measured, at least one of the approach groups differs significantly from the control group. For example, the percentage of black students is higher in Approach One than in the control group: 37.8% vs. 17.5% and lower in Approach Two and Three:

FIG. 5-1 CONTINUED

ITBS Testing in Spring 2000 and Spring 2001

Middle Schools	School	ITBS Reading			ITBS Language			ITBS Math		
Group		Grade								
		6	7	8	6	7	8	6	7	8
Pilot	Horace Mann	x	x		x	x		x	x	
	% Tested	93%	93%	8%	91%	93%	8%	91%	81%	8%
Control	Baker	x	x	x	x	x	x	x	x	x
	Cole	x	x	x	x	x	x	x	x	x
	Gove	x	x	x						
	Grant	x	x		x	x		x	x	
	Hamilton	x	x	x	x	x	x	x	x	x
	Hill	x	x	x	x	x	x	x	x	x
	Kepner	x	x	x	x	x	x	x	x	x
	Kunsmiller									
	Lake	x	x	x	x	x	x			
	Merrill	x	x	x	x	x	x	x	x	x
	Morey	x	x		x	x		x	x	
	Place	x	x	x	x	x	x	x	x	x
	Kishel	x	x		x	x		x	x	
	Skinner	x	x	x	x	x	x	x	x	x
	Smiley	x	x	x	x	x	x	x	x	x
	Henry	x	x	x	x	x	x	x	x	x
	Martin Luther King, Jr.	x	x	x	x	x	x	x	x	x
	Moore	x	x	x						
	% Tested	90%	90%	75%	87%	88%	76%	79%	78%	70%

"At Least 65% of Students Tested, by School, Subject, and Grade"

9.3% and 11.3% respectively (see Figure 5-2). Inequalities between the control and pilot approach groups exist for all measured factors except gender, disability, and grade retention.

The student, teacher, and school factors listed in Figure 5-2 were assessed for their impact on achievement and a common set of covariates was selected for use in all of the elementary school HLM models. Student level covariates include grade, test exclusion indicator, socioeconomic status (receives school lunch aid or non-aided), limited English proficiency (LAU=A or B, LAS=1 or 2), any disability, ethnicity (black, Hispanic, other), gender, and retained a grade. One teacher level covariate is included—an indicator of whether the teacher holds a Masters or Doctorate degree. School level covariates include the percent of students in the school who speak Spanish, the percent who speak a language other than Spanish or English, and total school enrollment.

Middle Schools

As with the elementary school sample, there are significant differences between the middle school pilot and control school students. As shown in Figure 5-3 the pilot school population is poorer, less likely to speak English, and was less likely to be performing at or above grade level before the pilot began than students in the control middle schools. While we can use multivariate techniques to take these differences into account, the dramatic difference between the pilot school (Horace Mann) and the rest of the district make it harder to generalize results from this one school to the whole district. A second middle school is currently participating in the pilot, so we will be able to generalize more accurately from the findings in subsequent reports.

Covariates selected for the middle school HLM analysis include grade, socioeconomic status, limited English proficiency, any disability, ethnicity (Hispanic, non-Hispanic), percent of students in the school who speak Spanish, percent of students who speak a language other than Spanish or English, percent of teachers with three or less years, four to 10 years, and more than 10 years experience, percent of teachers with a Masters degree, stability of student population (percent at school three years), and percent of teachers not fully licensed. Because we do not have teacher assignments for the controls, we cannot control for teacher characteristics.

E. Elementary School ITBS Results

Baseline NCE Scores and Change in NCE Scores

The elementary school students in both control and pilot groups were performing below grade level (average NCE < 50) at the start of the study, with the exception of Approach Two students in math and Approach Three students in reading. Excluding students who should probably not be tested, Approach Three students were also performing at grade level in math. (Figure 5-4)

Between the baseline and follow-up tests, students in the control group experienced statistically significant increases of just under one NCE in reading and language, and a decrease of just over one NCE in math. Approach One students achieved as expected—the average change in NCE scores on all three tests were not statistically different from zero, indicating the expected growth for a year in school. Approach Two and Three students also achieved the expected increase in reading and language, but saw statistically significant declines in math achievement of four NCEs and eight NCEs. While statistically significant, it is important to note both that these findings do not necessarily constitute a trend, and that the higher a school's starting place on an NCE scale, the greater the likelihood that that school's score will decline or stay the same instead of rising.

Elementary Results by Approach

In Figure 5-5, the mean changes in NCE scores reported in Figure 5-4 are tested to see if they differ from the control group. The only significant differences were lower achievement levels for Approach Three in reading and math, and for Approach Two in math.

Between School and Within School Differences

As noted in the discussion of methodology, hierarchical models are needed to account for correlation between students attending the same school and between students in the same classroom. Before constructing any models, we can use HLM to partition the variance in student growth between student factors and non-student factors (school and classroom). For reading, school and classroom factors account for 17% of the variation in student growth, leaving 83% to be explained by child-level characteristics. For language, 13% of the variation is attributable to school and classroom factors, and for math 26%.

When we reexamine the data treating school and teacher nested within school as random effects, only the negative effect of Approach Three on math achievement remains statistically significant. Approach Three students lost 5.8 NCE points while the control group lost only 1.9 NCE points. Excluding ineligible students (Figure 5-5 Non-Excluded Students) from the models produces the same results—no program effect for Approaches One and Two, and a roughly equivalent negative effect of Approach Three on math achievement.

FIG. 5-2

Elementary School Student Demographics

Demographic	Control		Approach One		Approach Two		Approach Three	
	Percent	N	Percent	N	Percent	N	Percent	N
Reading Sample	98.9	4909	87.9	886***	99.1	675	89.9	630***
Language Sample	75.7	3755	85.9	866***	36.4	248*	24.7	173***
Math Sample	73.8	3662	80.3	809***	74.6	508	48.4	339***
Grade								
2	16.8	836	21.7	159*	25.3	172*	19.8	139*
3	24.1	1194	24.0	180	34.2	233	28.5	200
4	28.8	1428	27.9	281	18.8	128	20.1	141
5	30.3	1504	26.4	266	21.7	148	31.5	221
Total	00.0	4962	100.0	886	100.0	675	100.0	630
Free or Reduced Lunch	61.0	3028	65.1	656*	72.1***		44.4	311***
Ethnicity								
Native American	1.4	69	1.1	11*	0.6	4*	0.7	5***
Black	17.5	870	37.8	381	9.3	63	11.3	79
Asian	3.8	190	4.0	40	2.9	20	4.1	29
Hispanic	42.8	2124	38.2	385	64.0	436	31.0	217
White	34.4	1709	18.9	191	23.2	158	52.9	371
Male	51.1	2536	49.1	495	52.4	357	51.6	362
Any Disability	11.7	578	10.5	106	12.0	82	9.3	65
Not Excluded from ITBS	81.1	4022	82.8	835	85.2	580*	77.7	545*
Retained a Grade	0.9	46	0.6	6	0.7	5	0.3	2
LAS								
None	73.6	3654	73.9	745	77.0	524*	72.6	509
1	0.7	36	1.2	12	2.2	15	3.4	24
2	1.0	51	1.0	10	1.8	12	4.1	29
3	4.3	212	3.9	39	3.5	24	4.4	31
4	12.8	637	12.6	127	9.7	66	9.7	68
5	7.5	372	7.4	75	5.9	40	5.7	40
LAU								
No English	0.6	28	0.8	8	1.3	9*	3.3	23***
Some English	9.1	452	8.9	90	5.4	37	12.4	87
English & other Language	11.2	555	12.6	127	8.7	59	8.8	62
Mostly English	5.2	256	4.4	44	6.6	45	4.6	32
Non LAU	74.0	3671	73.3	739	78.0	531	70.9	497

FIG. 5-2 CONTINUED

Elementary School Student Demographics

Demographic	Control		Approach One		Approach Two		Approach Three	
	Percent	N	Percent	N	Percent	N	Percent	N
Teacher Changed 2000/2001	4.3	212	5.4	54	14.1	96	2.0	14**
Changed Schools	31.0	1539	3.3	33*	38.6	263*	19.5	137***
Baseline Reading Achievement Group								
Not Tested	0.7	33	2.0	20*	0.4	3	1.9	13.0***
Lowest	40.9	2027	38.0	378	42.0	286	28.1	197
Middle	28.7	1424	35.1	354	24.5	167	21.1	148
Highest	29.8	1478	25.4	256	33.0	225	48.9	343
Baseline Language Achievement Group								
Not Tested	10.3	513	6.6	66*	10.7	73	40.7	285***
Lowest	34.3	1704	35.5	358	36.4	248	22.8	160
Middle	28.7	1426	30.3	305	25.6	174	12.7	89
Highest	26.6	1319	27.7	279	27.3	186	23.8	167
Baseline Math Achievement Group								
Not Tested	16.7	828	10.1	102*	21.9	149*	45.9	322***
Lowest	30.5	1512	33.9	342	23.2	158	17.6	123
Middle	27.6	1371	31.9	321	28.6	195	15.1	106
Highest	25.2	1251	24.1	243	26.3	179	21.4	150
Class Size								
1 to 5	1.2	57	0.2	2*	1.0	7*	3.1	22***
6 to 15	12.1	600	3.4	34	12.5	85	16.1	113
16 to 25	57.1	2834	20.8	210	78.9	537	59.6	418
26 to 30	28.5	1412	75.6	762	7.6	52	21.1	148
More than 30	1.2	59	0.0	0	0.0	0	0.0	0
Taught by Licensed Teacher	74.2	3680	60.5	610*	69.3	472*	69.8	489*
Taught by Endorsed Teacher	86.5	4294	64.1	646*	81.6	556*	88.3	619*
Teacher Education Level								
Taught by Teacher with Advanced Degree	41.1	2038	38.3	386	46.3	315*	53.9	378***
Taught by Teacher No Advanced Degree	55.0	2730	44.8	452	48.5	330	38.7	271
Taught by Teacher/ No Degree Info	3.9	194	16.9	170	5.3	36	7.4	52

FIG.5-3

Middle School Student Demographics

Demographic	Control		Pilot	
	Percent	Number	Percent	Number
Grade				
6	50.2	3061	48.9	156
7	49.8	3036	51.1	163
Total		6097		319
Free or Reduced Lunch**	67.0	4086	91.2	291
Ethnicity**				
Native American	1.3	78	0.9	3
Black	24.1	1469	1.3	4
Asian	4.0	245	0.9	3
Hispanic	48.6	2964	90.0	287
White	22.0	1341	6.9	22
Any Disability**	11.7	712	10.7	34
Excluded from ITBS Testing	17.8	1085	34.5	110
Retained a Grade**	1.5	94	0.3	1
LAS				
None	65	3961	38.9	124
1	1.5	91	2.8	9
2	1.7	103	4.4	14
3	3.5	215	4.7	15
4	15.4	940	24.8	79
5	12.9	787	24.5	78
Language Proficiency**				
No English	1.5	91	3.8	12
Some English	9.2	563	23.5	75
English & other Language	16.0	978	21.3	68
Mostly English	6.5	397	8.5	27
Non LAU	66.7	4068	43.0	137
Male	50.2	3060	50.8	162
Baseline Achievement Group**				
Missing	0.6	36	0.3	1
Lowest	40.3	2459	60.5	193
Middle	29	1768	29.2	93
Highest	30.1	1834	10	32
Baseline >= Grade Level**	35.6	2168	15.4	49

Elementary Results by Baseline Student Achievement

As previously noted, ITBS allows for a direct comparison between a student's scores on a particular test in one year to that same student's score the next year, yielding a representation of student growth. This yields by far the most accurate representation of classroom change, as it implicitly controls for—more effectively than any explicit statistical program—much of the variability across students. This is a critical difference from analysis of current year scores, where the performance of each child has a significant effect on the classroom mean. In that instance, we need to rely on imperfectly measured demographics to separate differences explained by the pilot, from differences explained by baseline differences in the students attending pilot and control schools. Thus, while we may see no difference in student growth by SES in the ITBS analysis because SES is implicitly controlled for already, we could expect to see differences by SES in students' scores for one year in the CSAP analysis. Measures of student growth will ultimately be required for a successful pay for performance program.

In order to examine possible differences in the impact of the pilot on students of differing academic ability, we assigned students to three baseline achievement levels (low, middle, or high). Students scoring 39 NCE or below are the low achievement group, those scoring 40 to 56 were assigned to the middle group, and above 55 to the high group. Baseline language achievement groups were created in the same way using breakpoints at 37 and 55, while the math groups

FIG.5-4

Elementary School Mean Scores and Mean Change in Score by Approach, 2000 and 2001

Hierarchical Linear Models

	All Students						Non-Excluded Students					
	Reading		Language		Math		Reading		Language		Math	
	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
Spring 2000 ITBS NCE Scores												
Control	45.5	4909	43.9	3755	46.4	3662	49.5	3982	47.5	2984	49.7	2989
Approach One	46.1	886	46.2	886	44.8	809	49.4	732	49.2	708	47.0	674
Approach Two	46.3	675	35.8	248	50.0	508	48.8	576	37.0	191	51.7	433
Approach Three	57.8	630	34.9	173	48.2	339	64.1	512	44.1	81	57.1	218
Spring 2001 ITBS NCE Scores												
Control	46.5	4909	44.8	3755	45.1	3662	50.1	3982	7.6	2984	47.8	2989
Approach One	46.1	886	46.7	886	44.0	809	49.7	732	49.2	708	46.0	674
Approach Two	46.2	675	37.3	248	45.9	508	48.5	576	38.2	191	51.7	433
Approach Three	57.4	630	33.5	173	39.6	339	63.8	512	37.2	81	57.1	218
Change in ITBS NCE Scores Spring 2000 to Spring 2001												
Control	0.9***	4909	0.9***	3755	-1.3***	3662	0.6**	3982	0.1	2984	-1.9***	2989
Approach One	0.0	886	0.5	886	-0.8	809	0.4	732	0.0	708	-1.0	674
Approach Two	-0.1	675	1.6	248	-4.1***	508	-0.3	576	1.1	191	-4.0***	433
Approach Three	-0.3	630	-1.4	173	-8.6***	339	-0.2	512	-6.9***	81	-11.6***	218

Null Hypothesis: Mean Change in NCE Score = 0, Significance Levels Indicated by * >0.05, **<0.01, ***<0.001

FIG.5-5

HLM Models Predicting Mean Change in ITBS Scores Spring 2000 to Spring 2001

Test	All Students								Non-Excluded Students							
	Unadjusted Model				Adjusted Model (HLM)				Unadjusted Model				Adjusted Model (HLM)			
	Control	One	Two	Three	Control	One	Two	Three	Control	One	Two	Three	Control	One	Two	Three
Reading	0.9	0.0	-0.1	-0.3*	0.7	0.6	1.1	-0.3	0.6	0.4	-0.3	-0.2	0.5	0.4	1.0	-0.5
Language	0.9	0.5	1.6	-1.4	0.8	0.6	0.0	-0.7	0.1	0.0	1.1	-6.9***	0.0	0.0	-0.3	-3.3
Math	-1.3	-0.8	-4.1***	-8.6***	-1.9	-0.3	-2.0	-5.8**	-1.9	-1.0	-4.0*	-11.6***	-2.7	-0.8	-1.4	-7.2*

Unadjusted Models are linear regression models in which approach is the only covariate.
Adjusted Models treat pilot as fixed, school as random, and include the following covariates: grade, SES, limited English proficiency, any disability, Hispanic, black, male, retained a grade, excluded from ITBS testing, Class size < 15 students, % teachers with an advanced degree, % students who speak Spanish, % students who speak other than English or Spanish, and school enrollment. Significance Levels are indicated by * <0.05, ** < 0.01, *** < 0.001

used breakpoints of 36 and 55. The resulting achievement groups are roughly equal based on the scores of students eligible to take the ITBS.

The lowest achievers at baseline showed no significant differences between control and the three pilot approach groups in reading and language (Figure 5-6). In math, Approach Three performed significantly lower than the controls, with a decline of 0.2 NCE points compared to a control group increase of 5.0 NCE. The same is true of the middle achievers; the only significant difference between the pilot and controls was for Approach Three math. On average, the Approach Three middle achievers had a change in NCE score 9 points smaller than controls. The same finding is true in the high achievement group, where the Approach Three change in NCE scores was 8.5 points lower than controls on the math test. In addition, the Approach Three high achievers' NCE change was also significantly lower on the language test. However, once the models were adjusted for student and teacher characteristics, all of these differences between control and Approach Three students lost statistical significance.

A potential concern regarding pilot schools is that teachers in one approach might perform better than the others because pilot teachers would have an incentive to give more attention to the students with the best chance of improving. The analysis by achievement level indicates that this has not been the case.

Elementary Results by Socioeconomic Status

The only indicator of socioeconomic status available is whether a child receives free or reduced rate lunch. Aided and non-aided students, as we have identified these students in Figure 5-7, showed no statistically significant difference between control students and pilot students in reading and language. Approach Three students performed worse than control students in math (unadjusted declines 7.6 and 10 NCE points larger than the control's decline) whether they received lunch assistance or not. After adjusting for student and teacher characteristics, the effects were virtually identical—declines of 4.5 NCE points more than the control group decline in math scores but statistically significant for only the aided students. Non-aided Approach One students performed better than control students in math (an increase of 1.1 NCE points versus a decrease of 2.9 NCE points for the controls).

Elementary Results by Ethnicity

The Approach Three negative effect on math NCE scores is seen for Hispanic (decrease of 5.8 NCE points from the previous academic year), black (decrease of 3.2 NCE), and white students (decrease of 4.0 NCE). The effect is statistically significant before adjusting for covariates, but is significant only for Hispanic students in the adjusted model. There are no other differences between pilot and control groups. (Figure 5-8)

F. Middle School ITBS Results

Middle School Baseline NCE Scores and Change in NCE Scores

Similar to the elementary school students, middle school students were also performing below grade level at baseline in reading, language, and math (Fig. 5-9). After we remove the students who should be excluded from testing, the control group performed at grade level on the language test. As one would expect from the differences in student demographics, mean baseline NCE scores for all three tests are substantially lower for the pilot sample. The paired comparison of means test, which tests whether the change in NCE score for the pilot differs from zero, shows that the pilot school had statistically significant gains in reading (1.8 NCE points) and language (5.5 NCE points) as well as a gain of 1.2 NCE points in math which was too small to be statistically significant. The paired comparison of means for the control schools shows a small but statistically significant decrease in reading of –0.5 NCE and statistically significant increases of 1.1 NCEs in language and 0.8 NCE in math.

Middle School Baseline ITBS Scores and Change in ITBS Scores

Although the pilot school started out with lower baseline scores, its students improved more than students in most of the control schools during the 2000–2001 academic year in reading and language. (Fig. 5-10) In reading, the gain of 1.8 NCEs was better than the performance of 10 of the 17 control schools, five of which had statistically significant losses. In language, the gain of 5.5 NCEs is significantly greater than for all but two control schools. This is an impressive gain, as six of the controls also had statistically significant, but smaller, gains.

FIG.5-6

Change in ITBS Scores Spring 2000 to Spring 2001 By Baseline Achievement Level

Test	Unadjusted				Adjusted			
	Control	One	Two	Three	Control	One	Two	Three
Low Baseline Achievement Group								
Reading	5.702	6.283	4.692	3.972	5.420	5.884	5.186	5.011
Language	6.991	7.415	6.277	6.017	6.955	8.376	5.317	4.189
Math	5.039	6.079	4.687	-0.161*	5.023	6.411	5.648	1.481
Middle Baseline Achievement Group								
Reading	0.557	-0.418	-3.472	0.119	0.077	-0.424	-1.937	-0.163
Language	-0.019	-0.863	-2.416	-6.7071	0.281	-1.711	0.053	-1.590
Math	0.047	-0.929	-4.029	-9.192**	-1.050	-0.632	-2.256	-5.308
High Baseline Achievement Group								
Reading	-4.762	-5.896	-5.244	-3.047	-4.626	-4.968	-2.584	-3.964
Language	-7.819	-6.935	-13.843	-16.598*	-7.688	-7.335	-12.091	-10.377
Math	-9.084	-8.237	-13.139	-17.679**	-9.789	-8.227	-11.501	-11.779

Unadjusted Models are linear regression models in which approach is the only covariate.

Adjusted Models treat pilot as fixed, school as random, and include the following covariates: grade, SES, limited English proficiency, any disability, Hispanic, black, male, retained a grade, excluded from ITBS testing, Class size < 15 students, % teachers with an advanced degree. Significance Levels are indicated by * <0.05, ** < 0.01, *** < 0.001

Middle School HLM Models

Figure 5-11 presents the results of the hierarchical linear models for all students and for a number of sub-populations. In order to account for correlation between the students within each school, the models treat school as a random effect. The unadjusted models predict mean change in score and test for a difference between pilot and control schools. The adjusted models also treat school as a random effect, and adjust for child level characteristics (grade, socioeconomic status, limited English proficiency, any disability, and Hispanic/Non-Hispanic) and for school level characteristics (percent of teachers <=3 yrs experience, 4-10 yrs experience, 11+ yrs experience). We were unable to treat classroom as a random effect as we did for the elementary schools because we did not have student assignments by teacher. The predicted means for sub-populations are presented. However, the sample size for the middle schools is too small to make statistical inferences on sub-groups.

Correcting for the correlation among students in the same school, we found that the increases noted earlier in reading and language NCEs are still statistically significant when compared to the control schools. After adjustment for child and school characteristics, however, the differences between the pilot and control schools are not significant. Increases in pilot student reading and math NCEs occurred in the Hispanic, low achievement, aided and non-aided sub-populations and all of the subgroups in Figure 5-11 experienced an increase in language scores. In general, disadvantaged groups (Hispanic, low achievers, and aided students) have better performance in the pilot school.

G. Comparison of ITBS Results to Teacher Objectives

One way of testing the relationship of objectives to achievement is to compare the ITBS reading scores of students by the quality of their teacher's objectives according to the rubric presented in Chapter IV. The Spring 2000 mean reading NCE increased as the quality of the objective increased, from 41 for the lowest rubric score to 52 for the middle two levels to 71 for the highest score. Also, the percent of children performing at or above grade level at the end of the previous year increased in a similar manner. The percent at grade level ranges from 33% for the lowest quality, the mid-50s for the middle two groups, up to 84% for those with excellent objectives.

FIG.5-7

Elementary School Mean Change in NCE Scores by SES, Spring 2000 to Spring 2001

Test	Unadjusted				HLM (Adjusted)			
	Control	One	Two	Three	Control	One	Two	Three
Aided Students								
Reading	1.3	1.0	0.0	-0.3	1.0	1.0	1.6	0.0
Language	1.7	1.3	1.5	-0.8	1.6	1.8	0.9	-0.6
Math	-0.7	-0.8	-3.6	-8.2***	-1.2	-0.4	-1.3	-5.8*
Non-Aided Students								
Reading	1.098	-0.405	-1.532	-0.575	0.471	0.312	0.967	-1.080
Language	-0.497	-1.303	0.711	1.20	-0.638	-0.962	3.781	-2.185
Math	-0.990	-0.061	-5.465	-11.127***	-2.911	1.074	-2.958	-7.482

Unadjusted Models are linear regression models in which approach is the only covariate.
Adjusted Models treat pilot as fixed, school as random, and include the following covariates: grade, limited English proficiency, any disability, Hispanic, black, male, retained a grade, excluded from ITBS testing, Class size < 15 students, % teachers with an advanced degree. Significance Levels are indicated by * <0.05, ** < 0.01, *** < 0.001

FIG.5-8

HLM Models Predicting Mean Change in NCE by Ethnicity, Spring 2000 to Spring 2001

	Unadjusted				Adjusted			
Hispanic Students	**Control**	**One**	**Two**	**Three**	**Control**	**One**	**Two**	**Three**
Reading	1.8	1.0	0.4	-1.2	1.4	0.9	1.8	-1.2
Language	1.8	1.7	3.0	-0.4	1.9	1.7	2.6	-1.5
Math	-0.1	1.7	-3.6	-8.6	-0.7	1.7	-0.8	-6.6
Difference from Control		1.8	-3.5	-8.5***		2.4	0.0	-5.8**
Black Students								
Reading	-0.8	0.6	-0.6	0.1	-0.7	-0.9	0.1	2.2
Language	-0.4	-1.0	-3.2	-3.8	-0.7	0.0	-6.3	-4.3
Math	-0.3	-2.3	-4.3	-8.2*	-1.8	-0.9	-2.8	-4.9
White Students								
Reading	1.1	-0.2	-1.1	-0.1	0.7	0.5	0.8	-0.3
Language	-0.1	-0.6	-3.4		0.1	-1.0	-4.9	
Math	-2.2	-0.7	-5.3	-13.2***	-2.9	-1.9	-3.9	-6.9

Unadjusted Models are linear regression models in which approach is the only covariate.

Adjusted Models treat pilot as fixed, school as random, and include the following covariates: grade, SES, limited English proficiency, any disability, male, retained a grade, excluded from ITBS testing, Class size < 15 students, % teachers with an advanced degree, % students who speak Spanish, % students who speak other than English or Spanish, and school enrollment.

Significance Levels are indicated by * <0.05, ** < 0.01, *** < 0.001

This suggests that the teachers who are able to create high quality objectives are already teaching high-achieving students. This might be due to a concentration of more highly educated or experienced teachers in some schools, to pre-existing professional development programs in certain schools, or to the efforts of good teachers raising the level of their entire school.

To judge whether student performance during the 2000-2001 academic year is related to the quality of objectives, we subtracted the previous spring's reading NCE score from the Spring 2001 score. As previously noted, the NCE scores are norm-referenced and scaled such that a change of zero means that a child has achieved a year's expected development in reading relative to a representative national sample, given their baseline reading level. The mean change in student reading scores is close to zero and sometimes even negative for the Needs Improvement, Partial and Acceptable quality categories. However, the students of teachers with Excellent quality objectives show increases of 3.6 (Objective One) and 1.4 (Objective Two) NCE points on average. This indicates that the teachers who wrote excellent objectives achieved more than a year's development with their students, on average, while teachers with less than excellent objectives obtained only the expected year's development on average, and no more.

The relationship between whether objectives were met, quality of the objectives, and student achievement is less clear cut. We examined the mean change in reading NCE scores by quality of objective and whether the objective was met. For excellent objectives, it does not matter whether objectives were met or not met; the students experienced a 2.8 NCE increase if the

FIG.5-9

Middle School Mean Scores and Mean Change in Score, Spring 2000 and Spring 2001

School Group	All Students						Non-Excluded Students					
	Reading		Language		Math		Reading		Language		Math	
	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
Spring 2000 ITBS NCE Scores												
Control School	42.5	6061	46.7	5913	40.6	5304	46.9	4984	50.0	4847	43.7	4414
Pilot School	32.1	318	36.7	313	32.8	292	38.7	209	42.6	206	37.2	185
Spring 2001 ITBS NCE Scores												
Control School	42.0	6061	47.800	5913	41.4	5304	45.9	4984	50.8	4847	44.1	4414
Pilot School	33.9	318	42.100	313	34.0	292	39.5	209	47.1	206	38.2	185
Change in ITBS NCE Scores Spring 2000 to Spring 2001												
Control School	-0.5**	6061	1.1***	5913	0.8***	5304	-1.0***	4984	0.8***	4847	0.4***	4414
Pilot School	1.8*	328	5.5***	313	1.2	292	0.7	209	4.6***	206	1.0	185

Null Hypothesis: Mean Change in NCE Score = 0, Significance Levels Indicated by * >0.05, **<0.01, ***<0.001

teacher succeeded and a 2.4 point increase if the teacher failed to meet Objective One. However, when Objective One Needs Improvement or is Acceptable, unmet objectives are associated with decreases in reading NCEs of 1.2 and 1.5 points. When the first objective is Partially acceptable, unmet objectives are associated with a 2.2 increase while met objectives were associated with a slight decrease of –0.4 NCEs. Similarly, there are counter-intuitive results for Objective Two, where the Acceptable category dropped 1.5 NCEs when the teacher succeeded and increased slightly when the teacher failed. It appears, based on a year's evidence, that teachers who failed to meet the lower quality objectives had students showing lower achievement, and that excellent objectives are associated with improved student achievement regardless of whether the teacher succeeded in fulfilling the objective. Achievement levels for met and unmet middle quality objectives were mixed.

H. Summary

Utility of the ITBS

Because the ITBS is norm-referenced, developmentally scaled, and widely used within the Denver Public Schools, it is the most useful indicator of changes in student achievement. Because it allows us to focus on student growth, it effectively eliminates the statistical need to account for differences, such as race/ethnicity or socioeconomic status. Using scores from the same child is a strong control. This same advantage leads us to expect fewer statistically significant results than for an assessment such as CSAP, where it is not possible to compare the scores of students in one year against the same students the following year. At the same time, it leads to greater confidence that where there are statistically significant results they are also meaningful.

The weakness of ITBS is that it is not closely linked to classroom instruction in many instances. ITBS is most closely linked to elementary classroom instruction, but it is still a general achievement test. If Denver's classroom teachers are teaching reading and math according to the state and district standards, CSAP is likely to be more closely linked to what a teacher is actually teaching. When CSAP is scaled such that it can be used to measure student progress, it should be the strongest indicator yet of true student growth, as well as of the performance of classroom teachers.

FIG.5-10

Middle School Mean Change in ITBS NCE Scores by School, Spring 2000 to Spring 2001

	Reading			Language			Math		
	Mean Change in NCE	H0: Mean=0	H1: Pilot Mean= Control(i)	Mean Change in NCE	H0: Mean=0	H1: Pilot Mean= Control(i)	Mean Change in NCE	H0: Mean=0	H1: Pilot Mean= Control(i)
Horace Mann	1.8	*		5.5	***		1.2		
Baker	-1.8	*	***	0.3		***	0.6		
Cole	-2.7	***	***	0.9		***	-0.2		
Gove	-0.6		*	1.7				not tested	
Grant	1.5			6.6	***		0.7		
Hamilton	-1.9	***	***	0.0		***	2.2	***	
Hill	0.2			-0.2		***	0.8		
Kepner	0.4			2.9	***	**	2.9	***	
Kunsmiller	-3.8	*	**	0.9		*	0.0		
Lake	-0.9		**	2.7	***	**	-6.0		
Merrill	1.0			2.5	***	**	2.0	**	
Morey	-0.1			2.4	**	**	1.4		
Place	-0.1		*	0.4		***	-0.2		
Rishel	-1.2	*	***	1.9	**	***	0.6		
Skinner	-0.4		*	2.2	**	**	-0.4		
Smiley	0.3			1.1		***	1.2		
Henry	-1.4	**	***	0.0		***	-0.7		*
MLK	0.3			-1.9	***	***	-0.2		

Statistical Significance indicated by * < 0.05, ** < 0.10, ***<0.001

Elementary Schools

Student growth was not significantly different for the reading and language subtests between elementary pilot and control schools. On the math subtest, students in Approach Three schools performed at a higher level, but worse in terms of year-to-year growth than the other schools by a statistically significant amount. However, teachers in Approach Three schools did not write their objectives toward ITBS measures.

While it might be expected that the ITBS results for teachers whose objectives were based on criterion-referenced tests and professional development would not show an effect on ITBS, Approach One teachers did address ITBS in their objectives, and most were judged to have achieved those objectives. For this reason, we might have expected some effect for this group. This result may be explained in part by several factors:

- Many teachers are not classroom teachers. According to Denver officials, more than half of the teaching force is composed of special subject teachers, special education teachers, and specialists.
- The parameters for setting objectives allow teachers to exclude certain children who cannot fairly be measured (e.g. those who are often absent or who are identified by a special needs or other teacher). For this reason,

FIG.5-11

Middle School Students — Mean Change in NCE Spring 2000 to Spring 2001

Hierarchical Linear Models

Population	Reading				Language				Math			
	Unadjusted		Adjusted		Unadjusted		Adjusted		Unadjusted		Adjusted	
	Control	Pilot	Control	Pilot	Control	Pilot	Control	Pilot	Control	Pilot	Control	Pilot
All Students	-0.5	1.8*	-0.5	0.7	1.4	5.5*	1.3	3.5	0.8	1.2	0.7	0.8
Non Excluded	-1.0	0.7	-1.0	1.0	1.2	4.6	1.1	2.9	0.4	1.0	0.3	1.5
Hispanic Students	-0.1	2.4	-0.1	2.1	2.0	5.5	2.0	3.7	0.8	1.5	0.9	0.4
Non Hispanic	-0.8	-3.6	-0.7	-4.0	0.9	5.3	0.7	2.8	0.5	-1.1	0.6	0.2
Low Achievement Group	4.8	5.3	4.8	5.2	4.8	9.8	4.9	8.4	7.0	7.0	6.7	7.9
Middle Achievement Group	-3.0	-2.7	-2.8	0.1	1.5	1.4	1.4	0.8	-1.1	-1.3	-1.2	0.0
High Achievement Group	-3.4	-1.4	-3.3	-1.0	-1.5	1.2	-1.4	-0.6	-4.4	-6.7	-4.0	-2.1
Aided Students	-0.7	1.8	-0.6	1.0	1.4	5.6	1.4	3.6	0.8	1.3	0.9	-0.1
Non Aided Students	-0.3	1.5	-0.2	-6.1	0.0	1.5	-0.2	-0.3	1.0	3.5	0.7	1.1

Unadjusted Models treat pilot as a fixed effect, School as a random effect.
Adjusted Models treat pilot as fixed, school as random, and include the following covariates: grade, SES, limited English proficiency, any disability, Hispanic, % teachers <=3 yrs experience, 4-10 yrs experience, 11+ yrs experience.

students considered by the principal in determining whether an objective was met may be somewhat different from students identified in our analysis.

- Teachers have the flexibility to focus at least one objective on a certain segment of the class. This may also mean that students excluded in some objectives may be a significantly different group from those in our analysis.
- Since two years of data are not indicative of a trend, this may also be a one-year variation.

Middle Schools

Unlike the elementary schools, the middle school pilot shows improvement in reading and language scores, and better performance on the part of the pilot school compared to the control schools. The pilot school improved an average of 1.8 NCEs points in reading, 5.5 NCEs in language, and 1.2 NCEs in math. In contrast, the control schools lost 0.5 NCEs in reading and gained 1.1 in language and 0.8 in math. The analysis of the middle school data is hampered by the small sample size and lack of teacher assignments for the control students, but it is reasonable to conclude that the pilot had a positive effect overall, and probably had a positive effect on Hispanic students, aided students, and low achievers as well. It is worth noting that the pilot middle school, Horace Mann, focused on writing in the 2000-2001 year, including a popular professional development program in writing, as well as the use of the Six-Trait program, and that many of its objectives were focused in this area.

Links Between Objective Quality and Student Achievement

A statistically significant linkage was found between objective quality and student achievement. Students whose teachers had objectives rated as Excellent, improved 3.6 NCE points in reading while the growth of students whose teacher's objectives were less than Excellent was close to zero (a year's expected growth). The main difference between an objective rated Excellent and one rated Acceptable according to CTAC's rubric is most often the specific identification of student learning on which the student results were to be measured. This finding suggests that building the capacity of teachers to develop high quality learning objectives would be beneficial to students and teachers alike.

CHAPTER VI

Colorado Student Assessment Program and 6+1 Trait Writing

A. Introduction

As described in Chapter III, CSAP has grown in importance during the past few years, and is used by the Colorado Department of Education to rank schools. The school-by-school results are published in the newspapers and announced on television. In the first year of the pilot, these rankings were described by the grading letters A-F. In the second year, the terminology changed such that low performing schools are rated "unsatisfactory" or "low." Colorado is like most states in that it does not, as yet, adjust its statistics for differences among children and classrooms, with the result that schools where the children primarily come from low-income families, are members of minority groups, or are non-native speakers of English, are generally ranked lower than other schools. As the state's largest urban school system, Denver consequently has a disproportionate share of schools with these ratings.

Colorado's CSAP is like many other state tests in another respect as well: it is not yet given according to a consistent plan. The grade levels at which the tests are given, and the content of those tests, have varied from year to year. For example, the reading test is now given to all students in grades three through 11, but previously it was not given in grades two, five, and eight. This prevents comparisons from grade two to grade three and, for the present, there is no comparison for grade four. New tests at new grades are being introduced, but because they are new they have no track record and may not be comparable. In this chapter, we focus on reading at the third and fourth grade level, but these results do not show individual student growth—the nature of the tests prevents using the third grade test as a starting place for measuring the growth of fourth graders. The Colorado Department of Education is working towards vertically scaling CSAP tests for that purpose.

CSAP does present one testing advantage over ITBS—it is based on Colorado State standards. To the extent that the curriculum and instruction at a given grade level addresses those standards, CSAP should provide a closer link to teacher performance than a more general test.

The second part of this chapter considers the 6+1 Trait Writing assessment (Six-Trait). This assessment differs substantially from either CSAP or ITBS, providing a better classroom tool for instruction in writing, but presenting difficulties in implementation and in use as a broad scale evaluative measure. It is an instructional program, rather than simply an assessment, which is based on teachers using a writing rubric to teach and assess student writing. Because of these differences, it also requires a different form of analysis.

B. Colorado Student Assessment Program (CSAP)

Research Approach

The differences between CSAP and ITBS have led to significantly different analyses. ITBS allowed for measuring student growth from one year to the next, which eliminates the need to adjust for many child level characteristics. Using the same child as his or her own control—having multiple years of data on a child—provides a strong control, as most child and family characteristics change little from one test administration to the next. Lacking the appropriate data to directly account for child level differences, our analysis of CSAP has focused on accounting for child level factors through statistical methods, and comparing these to classroom level factors, to test the extent to which these factors correlate with assessment outcomes.

For CSAP, we conducted quantitative analyses to:

- Compare the test scores of students in pilot classrooms to students in control classrooms.
- Determine if the quality of teachers' objectives or whether teachers achieved their yearly objectives is associated with their students' CSAP test scores.
- Describe the quantitative relationship, where one exists, between the quality of teachers' objectives, whether teachers achieved their objectives, the type of approach they used, and CSAP test scores of the students in their classrooms.
- Take into account additional extraneous variables such as gender, ethnicity, English proficiency, socioeconomic status, teacher education level, and teacher years of experience.

The study employed two different sets of analyses to address the research questions.

Between-Group Analysis: Method and Sample

To examine differences between groups, we used regression analysis. This allowed us to compare children in pilot classrooms to children in control classrooms. The test scores of students in pilot classrooms during the 1999-2000 and 2000-2001 school years were compared to test scores of students in the control schools. We examined the total reading, writing, and math (fifth grade 2000-2001 only) scale scores and the reading and writing standards scale scores of students by grade where possible. The Colorado Model Content Standards for reading and writing assessed by the CSAP include:

1. Students read and understand a variety of materials.
2. Students write and speak for a variety of purposes and audiences.
3. Students write and speak using conventional grammar, usage, sentence structure, punctuation, capitalization, and spelling.

4. Students apply thinking skills to their reading, writing, and speaking, listening, and viewing.
5. Students read to locate, select, and make use of relevant information from a variety of media, reference, and technological sources.
6. Students read and recognize literature as a record of human experience.

Figure 6-1 shows the measures of child outcomes that were examined by grade and by year. Sample sizes varied across grade, test and year; these figures can be found in the descriptive tables throughout this chapter.

Pilot Student Outcomes: Method and Sample

Our analyses of both ITBS and CSAP included using hierarchical linear modeling, a multilevel analysis method (Bryk & Raudenbush, 1992)[1]. HLM is used because it allows us to partition the variation in children's scores into between-child variance (to be explained by child level characteristics) and between-class variance (to be explained by classroom level characteristics) and ensures an accurate estimation of the effects. Our analysis fits into one of the most common multilevel analysis models, two-level school effects models, which allow us to examine child outcomes as a function of both child (Level 1) and classroom/teacher (Level 2) predictors. School effects models are designed for data on individuals nested within naturally occurring hierarchies (e.g. students within classes) (Singer, 1998)[2]. While the factors considered in conducting HLM on CSAP are different than those used for ITBS, as described above, the purposes and methodology are similar.

HLM sample size limitations required us to combine data across two years (1999-2000 and 2000-2001) for pilot school teachers and students in the third and fourth grades. This was possible because Colorado Department of Education converted scores on the prior year's reading tests to the same scale as the 2000-2001 scores. Our final HLM sample consisted of 92 third grade pilot classrooms with a total of 1,563 students and 97 fourth grade pilot classrooms with a total of 1,394 students (see Figure 6-2 for student background characteristics).

Differences in CSAP scoring from one year to the next limited us to predicting only the total reading scale scores for third and fourth grade students taking the English version of the CSAP. We were not able to conduct HLM analyses on any of the fourth grade reading or writing standards individually because of differences in scoring across years, and could not conduct HLM analyses on the fifth grade total reading and math scale scores due limitations in sample size at the classroom/teacher level (Level 2).

The analysis of our research questions is based on the school effects multilevel model shown in Figure 6-3. Scores for third and fourth grade pilot students on the 1999-2000 (administered Spring 2000) and 2000-2001 (administered Spring 2001) CSAP reading tests are the major dependent variables in our two-level models. We included three independent measures (predictors) in our between-classroom (Level 2) analyses, which estimated the effect of the quality of teachers' objectives, the approach used, and whether they met their objectives on students' CSAP reading scale scores. While estimating the effects of our predictors on student reading

FIG. 6-1

CSAP Content Areas and Standards Analyzed by Grade and Year

1999-2000		2000-2001		
3rd Grade	**4th Grade**	**3rd Grade**	**4th Grade**	**5th Grade**
Total Reading	Total Reading	Total Reading	Total Reading	Total Reading
	Reading Standards 1, 4, 5 & 6		Reading Standards 1, 4, 5 & 6	Total Math
	Total Writing		Total Writing	
	Writing Standards 2 & 3			

FIG. 6-2

Pilot Student Background Characteristics by Year

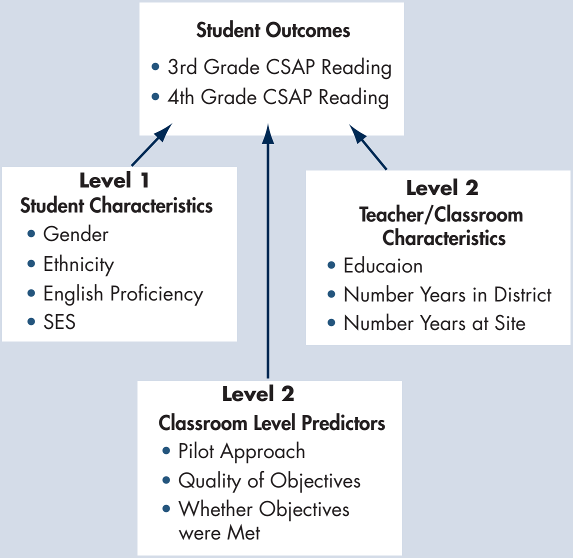
	3rd Grade		4th Grade	
	1999-2000	2000-2001	1999-2000	2000-2001
Gender				
Male	50.8%	49.6%	52.2%	50.5%
Female	49.2%	50.4%	47.8%	49.5%
Ethnicity				
Hispanic	43.7%	49.5%	44.9%	47.3%
White	30.9%	28.3%	30.4%	27.9%
Black	19.6%	18.8%	19.6%	19.7%
Other	5.8%	3.4%	5.1%	5.1%
English Proficiency				
No English	1.2%	5.6%	1.1%	2.5%
Some English, Mostly Another Language	16.6%	7.6%	22.7%	10.8%
English & Another Language Equally/ Speaks Mostly English	15.9%	15.3%	14.6%	15.0%
English Fluent	66.2%	71.5%	61.6%	71.7%
Socioeconomic Status				
Low (Free/Reduced Lunch)	57.5%	60.9%	59.3%	62.2%
High (Pay for Lunch)	42.5%	39.1%	40.7%	37.8%

outcomes, we took into account both students' socioeconomic backgrounds (Level 1) and also several characteristics of the teachers (Level 2). Because of the amount of missing data for both students and teachers, the control variables or factors in our multilevel models were limited to those that had sufficient data for analysis.

As Figure 6-3 shows, the child level factors considered in the HLM models included: English proficiency (LAU) which is scaled from 0 (No English) to 3 (Fluent in English), and three dummy coded (0,1) indicators of gender (male, female), ethnicity (Hispanic, Other), and socioeconomic status—SES (free/reduced lunch, pays for lunch). The classroom level control variables included three scaled variables: education level (BA, BA+, MA, MA+, Ed.D), number of years teaching in the district and at the site; and two dummy coded indicators of year (1999-2000, 2000-2001) and education (graduate degree, undergraduate degree). The independent variables, which test the outcomes of the pilot, consisted of:

1. Teacher approach used to obtain objectives:
   - Approach One
   - Approach Two
   - Approach Three
2. Whether or not teachers met one or both objectives during the school year.
3. The quality of teachers' objectives according to the rubric described in Chapter IV:
   - Level 1: Needs Improvement
   - Level 2: Partial
   - Level 3: Acceptable
   - Level 4: Excellent

FIG. 6-3



Results

Between-Group Analysis

An examination of group means for the 1999-2000 school year showed that statistically significant between-group differences only exist for students taking the English version of the CSAP and only on two outcomes. Pilot students performed significantly higher in third grade total reading (*t-statistic*=2.58, *p*=. 01), with a mean of 534.98 for pilot students compared to 527.06 for control students, and in fourth grade reading for Standard 1 (*t-statistic*=2.61, *p*=. 009), with a mean of 567.33 for pilot students compared to a mean of 557.50 for students in control classrooms. Figure 6-4 shows that although students who were in pilot classrooms had average scale scores that were somewhat higher on all dimensions, the differences were only big enough to be significant for two scores. For students given the Spanish version of the CSAP, Figure 6-5 shows that children in pilot classrooms had scale scores that were lower than control students' scale scores on almost every outcome and in two cases—fourth grade reading for Standards 4 and 6—control students actually had scale scores that were significantly higher than pilot students.

Simple linear regression analysis on test scores for the 2000-2001 school year revealed several statistically significant differences between pilot and control students who were given the English version of the CSAP.

Figure 6-6 shows that on all test scores but one (fourth grade reading Standard 5), students in pilot classrooms had scale scores that were significantly higher than control students on average. The mean of 536.65 for third grade pilot students in total reading was found to be significantly higher than the mean of 529.06 for control students (*t-statistic*=2.36, *p*=. 018). The differences between groups for fourth grade total reading and Standards 1, 4, and 6 were all statistically significant at the 0.0001 level of significance, and at the 0.01 level of significance for the fourth grade total writing scale scores. Pilot students who were in the fifth grade also performed significantly higher than control students in reading (*t-statistic*=2.53, *p*=. 011) and math (*t-statistic*=1.96, *p*=. 05), with a mean of 583.08 versus 575.74 for total reading and a mean of 463.19 versus 458.06 for total math.

Sample size requirements for any inferential statistical tests prevented us from performing between-group analyses on most of the Spanish-speaking student outcomes for 2000-2001. Figure 6-7 shows the descriptive statistics for these students by group. The sample size requirements were met for third grade students where control students performed higher than pilot students with a mean of 514.38 versus 507.26. This difference was not large enough to be statistically significant, probably due to the sample size.

Hierarchical Linear Modeling Analysis Results

We began the HLM analysis of CSAP results in the pilot schools by partitioning the variance in children's CSAP reading scores into between-child variance (to be explained by child level characteristics) and between-class variance (to be explained by classroom level characteristics). Only the variance in third and fourth grade outcomes that is between classrooms can be influenced by classroom/teacher factors.

To partition the amount of variation that can be explained by classroom level and between-classroom differences, we ran fully unconditional "null" HLM models first. These unconditional "null" models include the intercept (reading outcome variable) only and no predictors at either

FIG. 6-4

Reading and Writing Scale Score Descriptives by Group, 1999-2000

	Mean	SD	Minimum	Maximum	'n'
**3rd Grade Reading (Total/Standard 1)					
Pilot	534.98	76.81	294	756	812
Control	527.06	72.27	229	756	1973
4th Grade Total Reading					
Pilot	557.11	74.73	285	764	848
Control	552.39	73.84	285	799	1870
**4th Grade Reading (Standard 1)					
Pilot	567.33	98.10	285	895	848
Control	557.50	87.66	285	895	1870
4th Grade Reading (Standard 4)					
Pilot	551.33	65.15	300	720	851
Control	548.24	90.86	285	895	1870
4th Grade Reading (Standard 5)					
Pilot	566.82	97.80	285	895	848
Control	563.76	99.91	285	895	1870
4th Grade Reading (Standard 6)					
Pilot	560.38	96.29	285	895	848
Control	555.17	95.76	285	895	1870
4th Grade Total Writing					
Pilot	490.90	45.00	339	610	813
Control	488.76	44.00	300	720	1803
4th Grade Writing (Standard 2)					
Pilot	493.83	51.93	300	720	813
Control	491.18	51.40	300	720	1803
4th Grade Writing (Standard 3)					
Pilot	491.71	51.41	300	720	813
Control	490.00	51.41	300	720	1803

(Significance level indicated where applicable *p<. 05; **p<. 01; ***p<. 001)

Level 1 or Level 2. The results of these models and our final full HLM models, which include child level control variables (Level 1), classroom level control variables (Level 2), and significant predictors (Level 2), are presented in Figure 6-8 for analyses predicting third grade pilot student scores and Figure 6-10 for analyses predicting fourth grade pilot student scores.

CSAP reading scores for third grade pilot students: While ITBS allows the measurement of student growth using the same students from one year to the next, CSAP requires analysis at the individual grade level, and the use of statistical methods to control for child level variables.

Grade three CSAP is a test of basic reading proficiency. Using Equation 1 below (computed

FIG. 6-5

Lectura and Escritura Scale Score Descriptives by Group, 1999-2000

	Mean	SD	Minimum	Maximum	'n'
3rd Grade Reading (Standard 1)					
Pilot	503.00	49.13	348.00	587.00	68
Control	507.83	42.94	338.00	620.00	213
4th Grade Reading (Total Scale Score)					
Pilot	496.21	40.87	401.00	572.00	28
Control	506.84	39.16	368.00	582.00	120
4th Grade Reading (Standard 1)					
Pilot	503.75	51.71	425.00	633.00	28
Control	509.08	50.51	300.00	624.00	120
*4th Grade Reading (Standard 4)					
Pilot	465.11	87.26	300.00	574.00	28
Control	499.88	63.04	300.00	596.00	120
4th Grade Reading (Standard 5)					
Pilot	502.43	38.28	399.00	575.00	28
Control	502.17	52.65	300.00	644.00	120
*4th Grade Reading (Standard 6)					
Pilot	474.96	73.83	300.00	573.00	28
Control	501.67	56.19	300.00	586.00	120
4th Grade Writing (Total Scale Score)					
Pilot	497.96	37.92	416	582	28
Control	502.26	43.35	354.00	589.00	119
4th Grade Writing (Standard 2)					
Pilot	498.54	38.57	402.00	569.00	28
Control	507.61	51.47	300.00	627.00	119
4th Grade Writing (Standard 3)					
Pilot	497.93	43.98	425.00	624.00	28
Control	499.16	48.99	335.00	636.00	119

(Significance level indicated where applicable, but in favor of control students *p<.05; **p<.01; ***p<.001)

FIG. 6-6

Reading & Writing Scale Score Descriptives by Group, 2000-2001

	Mean	SD	Minimum	Maximum	'n'
*3rd Grade Reading (Total/Standard 1)					
Pilot	536.65	79.66	150	756	818
Control	529.06	77.29	150	756	2078
***4th Grade Reading (Total Scale Score)					
Pilot	561.79	111.70	180	940	853
Control	550.31	113.09	180	940	2139
***4th Grade Reading (Standard 1)					
Pilot	580.39	120.27	180	940	824
Control	563.43	107.96	180	940	2064
***4th Grade Reading (Standard 4)					
Pilot	555.52	111.70	180	940	824
Control	538.95	113.02	180	940	2064
4th Grade Reading (Standard 5)					
Pilot	566.80	106.30	180	940	824
Control	558.04	112.89	180	940	2064
***4th Grade Reading (Standard 6)					
Pilot	562.04	101.44	180	940	824
Control	546.39	96.35	180	940	2064
**4th Grade Writing (Total Scale Score)					
Pilot	492.62	40.02	312	720	796
Control	488.12	39.47	331	645	2001
*5th Grade Reading					
Pilot	583.08	69.89	220	803	817
Control	575.74	69.54	220	802	1990
*5th Grade Math					
Pilot	463.19	62.00	200	640	821
Control	458.06	63.81	200	682	2006

(Significance level indicated where applicable*p<. 05; **p<. 01; ***p<. 001)

FIG. 6-7

Lectura and Escritura Scale Score Descriptives by Group, 2000-2001

	Mean	SD	Minimum	Maximum	'n'
3rd Grade Reading (Total)					
Pilot	507.26	35.43	419.00	583.00	39
Control	514.38	42.33	348.00	599.00	182
4th Grade Total Reading					
Pilot	503.63	42.85	388.00	549.00	24
Control	504.31	44.12	300.00	606.00	129
4th Grade Reading (Standard 1)					
Pilot	505.38	43.13	399	566	24
Control	509.77	47.93	300	629	129
4th Grade Reading (Standard 4)					
Pilot	497.46	65.17	300	580	24
Control	493.22	64.26	300	620	129
4th Grade Reading (Standard 5)					
Pilot	500.96	44.34	401	583	24
Control	495.24	62.50	300	631	129
4th Grade Reading (Standard 6)					
Pilot	501.24	57.83	300	588	24
Control	496.86	71.40	300	641	129
4th Grade Total Writing					
Pilot	507.50	49.57	413.00	583.00	22
Control	504.85	33.43	374.00	581.00	122

using statistics from Figure 6-8), we estimated the intraclass correlation, which tells what portion of the total 100% of variation to be explained in third grade total reading CSAP scale scores occurs between classrooms. The estimated value of σ^2(4269.21) is nearly three times the size of the variance component between classrooms (τ_{00}=1472.46), which shows that although classrooms do differ in their average CSAP reading scores, there is even more variation among students within classrooms. The estimated variance components show that 74% of the total variance can be attributed to child differences, leaving 26% to be explained by classroom level differences.

Equation 1

$$\frac{\tau_{00}}{\tau_{00} + \sigma^2} = \frac{1472.46}{1472.46 + 4269.21} = 0.256$$

Next, we tested each child level factor in the HLM model to examine what child level characteristics were significant and should therefore remain in the model. Child gender and English proficiency were found to be significant positive child level (Level 1) factors (control variables) with female students having scale scores that were 8.67 points higher than males on average, and a difference of 16.38 scale score points associated with a one point increase in the English proficiency scale, on average. Socioeconomic

FIG. 6-8

Parameter Estimates and p-values for Final HLM Models Predicting Third Grade Reading

	Model 1 **γ (p-value)**	**Model 2** **γ (p-value)**	**Model 3** **γ (p-value)**	**Model 4** **γ (p-value)**
	Null Model	**Level 1** **Controls:* Gender, **LAU, SES	**Level 2** *Control:* # Years in District	**Level 2** *Predictor:* Approach
Fixed Effects				
LEVEL 1				
G_{00}—Total Reading *(Intercept-Classroom Mean)*	535.7(.001)	553 (.001)	552.4 (.001)	544.2 (.001)
G_{10}—Gender	—	8.67 (.004)	8.83 (.004)	8.87 (.004)
G_{30}—LAU	—	16.38 (.001)	16.03 (.001)	15.8 (.001)
G_{40}—SES	—	-31.9 (.001)	-31.6 (.001)	-31.7 (.001)
LEVEL 2				
Yrs in District	—	—	.88 (.005)	.90 (.004)
PREDICTORS				
Approach—Criterion Referenced	—	—	—	9.98 (.169)
Approach—Professional Development	—	—	—	16.4 (.063)
Random Effects				

σ^2 (R) (measurement error)	SD=65.34
	VAR=4269.2
		SD=61.8
		VAR=3819.5
			SD=61.79
			VAR=3818.6
				SD=61.8
				VAR=3816.3

$U_0/\ \tau_{00}$ (Variance of Classroom Mean)	1473 (.001)	718.2 (.001)	647.7 (.001)	593.5 (.001)
U_3/τ_{03} (Variance of LAU Slope)	—	304.3 (.001)	301.5 (.001)	312.9 (.001)
Level-1 Reliability Estimates				
B_0 (Intercept–Classroom Mean)	.779	.651	.631	.615
n	92	81	81	81
B_3 (LAU Slope)	—	.364	.362	.369
Deviance	17661.58	17490.91	17484.06	17480.27

Note: All variables that were not dichotomous were centered around the grand mean

* Gender (Male=0, Female=1); LAU (Scale of English Proficiency ranging from 0-3); SES (Free/Reduced Lunch=1, Pay for lunch=0);

** Significant random effects indicate that the "English proficiency (LAU)" varies significantly across classrooms; therefore, it was left to vary (variance component was not removed);

*** Approach (Criterion Referenced=1; Other Approach=0); Approach (Professional Development=1; Other=0) status (SES) was also found to be a significant child level factor, but the effect was negative. The negative coefficient indicated that children who were of lower socioeconomic status (free/reduced lunch) scored 31.94 points lower on average in reading than children who were of higher socioeconomic status (pay for lunch). Although ethnicity was initially found to be a significant factor, with Hispanic children scoring 7.54 points lower than other children, the effect was no longer significant when English proficiency and SES were entered into the model, therefore it was not included in our final models. Together, child gender, English proficiency, and SES explain 10% of the original 74% of variation in third grade students reading scale scores that is attributable to child differences (see Equation 2). The strongest child level predictor of third grade CSAP reading scores is SES.

Equation 2 (computed using statistics from Fig. 6-8)

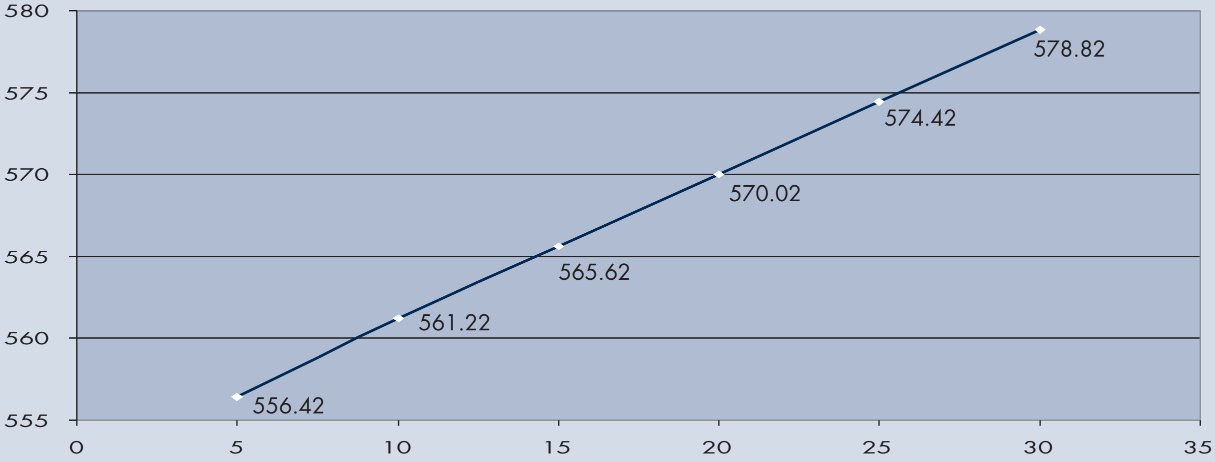
$$\frac{\sigma^2\,\textit{(original/null) - R}}{\sigma^2\textit{(original/null)}} = \frac{4269.21 - 3819.48}{4269.21} = 0.104$$

After the significant child factors were entered into the model, classroom factors (control variables and predictors) were entered to examine the extent to which we can predict third grade students' CSAP reading scale scores based on teacher characteristics, the approach they used, the quality of their objectives, and whether they achieved their objectives. We found the number of years a teacher has been in the district to be the only significant classroom level control variable. It is positively associated with students' CSAP reading scale scores, with a one year increase in the number of years a teacher has been in the district associated with a 0.88 point difference in their students' CSAP reading scores on average (Figure 6-9). While this produces a statistically significant correlation, we cannot determine cause and effect. This result could be explained either as an effect of experienced teaching, or by the tendency for senior teachers to end up in higher performing schools, or both. Further analysis with additional years of data will help to define the nature of the effect.

When examining the classroom level predictors, we found that the average reading scale score of students in classrooms of teachers who used Approach One was 16.40 points lower than for students in classrooms of teachers who used Approach Three, with a mean of 544.20 versus 560.60 respectively, and 9.98 points lower than students in classrooms of teachers who used Approach Two (mean=554.18). Neither of these

FIG. 6-9

Predicting Third Grade Students' CSAP Reading Scale Scores Based on Number of Years Teacher Has Been in the District



Note: A one year increase in the number of years in the district is associated with a 0.88 point increase in reading scores on average

FIG. 6-10

Parameter Estimates and p-values for Final HLM Models Predicting Fourth Grade Reading

	Model 1 γ (p-value)	Model 2 γ (p-value)	Model 3 γ (p-value)	Model 4 γ (p-value)	Model 4 γ (p-value)
	Null Model	**Level 1** **Controls:* Gender, *Ethnicity,* **LAU, SES	**Level 2** *Control:* # Years in District	**Level 2** ****Predictor:* Approach	**Level 2** ****Predictor:* Quality of Objectives
Fixed Effects					
LEVEL 1					
G_{00}—Reading *(Intercept/Classroom Mean)*	557.9 (.001)	581.5 (.001)	580.7 (.001)	587.5 (.001)	580.8 (.001)
G_{10}—Gender	—	7.28 (.036)	7.40 (.033)	7.49 (.031)	7.47 (.031)
G_{20}—Ethnicity	—	-10.4 (.004)	-10.4 (.004)	-10.7 (.003)	-10.4 (.003)
G_{30}—SES	—	-29.9 (.001)	-29.2 (.001)	-29.1 (.001)	-28.9 (.001)
G_{40}—LAU	—	21.2 (.001)	20.64 (.001)	20.88 (.001)	20.55 (.001)
LEVEL 2					
Yrs in District	—	—	1.01 (.004)	1.25 (.001)	1.00 (.004)
PREDICTORS					
Approach-Norm Referenced	—	—	—	-19.4 (.034)	—
Approach-Criterion Referenced	—	—	—	-2.24 (.809)	—
Quality of Objectives	—	—	—	—	13.0 (.020)
Random Effects					
σ^2 (R) (measurement error)	SD=59.42 VAR=3530.4	SD=54.91 VAR=3015.5	SD=54.84 VAR=3007.7	SD=54.88 VAR=3011.3	SD=54.87 VAR=3010.5
U_0/τ_{00} (Variance of Classroom Mean)	1982 (.001)	915.0 (.001)	855.3 (.001)	776.0 (.001)	786.0 (.001)
U_4/τ_{03} (Variance of LAU Slope)	—	217.0 (.002)	226.6 (.002)	223.4 (.002)	223.4 (.002)
Level-1 Reliability Estimates					
B0 (Intercept–Classroom Mean)	.825	.695	.684	.666	.669
n	97	80	80	80	80
B_3 (LAU Slope)	—	.304	.312	.309	.309
Deviance	15539.56	15308.98	15302.43	15296.94	15297.40

Note: All variables that were not dichotomous were centered around the grand mean

* Gender (Male=0, Female=1); Ethnicity (Hispanic=1, Not Hispanic/Other=0); LAU (Scale of English Proficiency ranging from 0-3); SES (Free/Reduced Lunch=1, Pay for lunch=0);

** Significant random effects indicate that the "English proficiency (LAU)" varies significantly across classrooms; therefore, it was left to vary (variance component was not removed);

*** Approach 1(Norm Referenced=1; Other Approach=0); Approach 2 (Criterion Referenced=1; Other=0);
Quality of Objectives is the average of the ratings of the quality of objectives 1 and 2 using a 1-4 scale

differences was statistically significant. The quality of teachers' objectives and whether teachers met their objectives were not found to be significant predictors of third grade students' reading scale scores.

CSAP reading scores for fourth grade pilot students: As Equation 3 (computed using statistics from Figure 6-10) shows the estimated intraclass correlation—or the portion of the total variation to be explained in fourth grade total reading CSAP scale scores that occurs between classrooms—is 36%. The estimated value of σ^2(3530.39) is nearly two times the size of the variance component between classrooms (τ_{00}=1981.65), which shows that although classrooms do differ in their average CSAP reading scores, there is even more variation among students within classrooms. The estimated variance components show that 64% can be attributed to child differences, leaving a significant amount, 36%, to be explained by classroom level differences.

Equation 3

$$\frac{\tau_{00}}{\tau_{00} + \sigma^2} = \frac{1981.65}{1981.65 + 3530.39} = 0.359$$

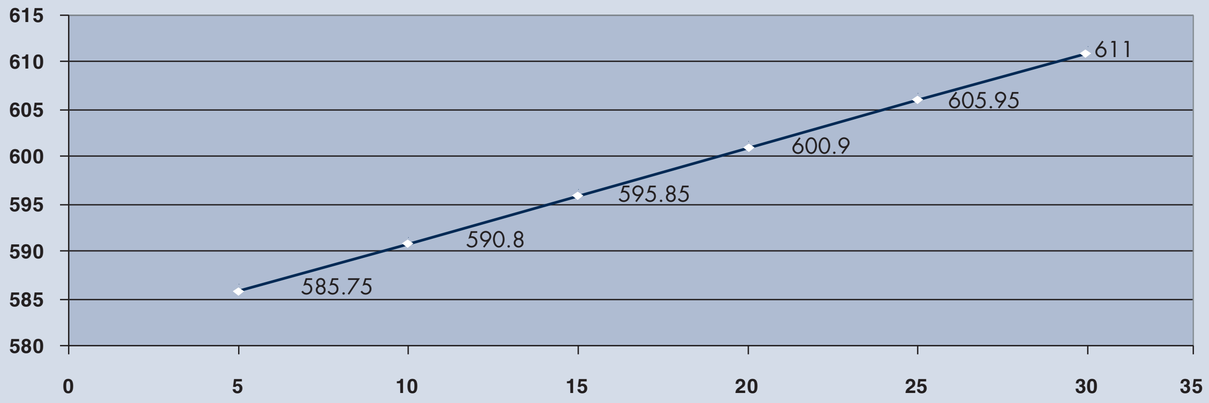
When the child level control variables were entered into the HLM models, the results for fourth grade students were similar to the results for third grade students. Child gender and English proficiency were found to be significant positive child factors, with female students having scale scores that were 7.28 points higher on average and a difference of 21.17 points associated with a one point increase in the English proficiency scale. Ethnicity (Hispanic or not) and SES were also found to be significant negative child factors for fourth grade students. Hispanic students had scale scores that were 10.36 points lower on average in reading than other students in their classrooms and children who were of lower SES (free/reduced lunch) scored almost 30 points lower on average in reading than children who were of higher SES (pay for lunch). Together, child gender, English proficiency, ethnicity, and SES explain 15% of the original 64% of variation in fourth grade students' reading scale scores that is attributable to child differences (see equation 4 below). The strongest child level predictor of fourth grade CSAP reading scores is SES.

Equation 4 (computed using statistics from Fig. 6-10)

$$\frac{\sigma^2\ (original/null)\ \text{-}\ R}{\sigma^2 (original/null)} = \frac{3530.39 - 3015.53}{3530.39} = 0.146$$

FIG. 6-11

Relationship Between Fourth Grade Students' CSAP Reading Scale Scores and Number of Years Teacher Has Been in the District



Note: A one year increase in the number of years in the district is associated with a 1.01 point increase in reading scores on average

After the significant child level variables were entered into the model, classroom level control variables and predictors were entered to examine the extent to which we can predict fourth grade students' reading scores based on teacher characteristics, the approach they used, the quality of their objectives, and whether they achieved their objectives. Similar to grade three, we found the number of years a teacher has been in the district to be a significant positive teacher level control variable for fourth grade classrooms. As Figure 6-11 shows, a one year increase in the number of years a teacher has been in the district is associated with a 1.01 point difference in their students' CSAP reading scores on average. As with third grade, cause and effect cannot be determined from these results. It is not clear whether teacher experience drives these higher results, or whether schools with higher results attract teachers with greater experience.

Differences between approaches: We also found the approach teachers used and the quality of their objectives to be significant classroom level predictors of fourth grade students' CSAP reading scale scores. Figure 6-12 shows the difference in students' average scale scores based on pilot approach. Teachers in Approach One schools had students with average scale scores that were 19.39 points lower than teachers in Approach Three schools, and 17.15 points lower than Approach Two teachers. The difference in the average scale scores was statistically significant for teachers in Approach One when compared to teachers in Approach Two (t=2.30, p=. 021) and Approach Three (t=2.12, p=. 034). Although the scale scores for students in Approach Two schools were 2.24 points lower on average than in Approach Three schools, the difference between these two approaches was not large enough to be statistically significant.

HLM analysis also indicated a significant positive relationship (p=. 02) between the quality of a teacher's objective, no matter the approach, and students' test scores. Specifically, a one point increase in the quality of a teacher's objective is associated with a 13 point increase in students reading scale scores on average.

To examine the amount of variation explained by the classroom factors, we used Equations 5 and 6. We are explaining 61% of the original 36% of variation in fourth grade students reading scale scores attributed to classroom level differences by approach (Equation 5) and we are explaining 60% by the quality of teachers' objectives (Equation 6).

Equation 5

$$\frac{\tau_{00}\,(original/null)\ \text{-}\ U_0}{\tau_{00}\,(original/null)} = \frac{1981.65 - 776.04}{1981.65} = 0.608$$

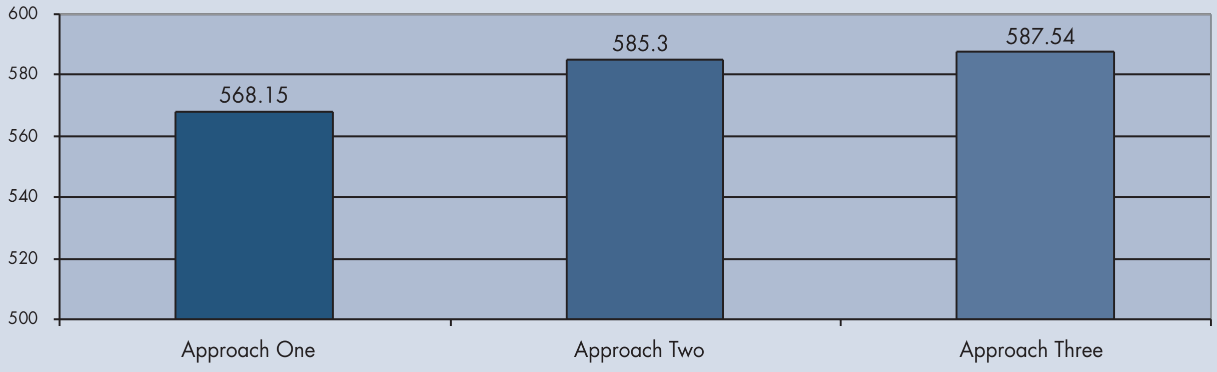
Equation 6

$$\frac{\tau_{00}\,(original/null)\ \text{-}\ U_0}{\tau_{00}\,(original/null)} = \frac{1981.65 - 786.03}{1981.65} = 0.60$$

Explaining more than half of the classroom level variation in third and fourth grade students' test scores is a significant finding. This finding indicates that the quality of the objectives, no

FIG. 6-12

Average Reading Scale Scores of Fourth Grade Students Based on Teacher Approach



matter what the approach, is significantly correlated with CSAP performance. The meaning of the approach effect is less clear. Partly due to the fact that schools self-selected into the three approaches, the demographics of the groups were not equal (Figure 5-2). As reported in Figure 5-4 of Chapter V, this resulted in the professional development schools starting with higher baseline reading ITBS scores than the other two approach groups and higher math ITBS scores than the norm-referenced group. A limitation of the CSAP test is that the grade three exam is too different from the grade four exam to be comparable, so we were not able to use the previous year's test score as a covariate. Thus we are not able to discern whether the approach effect is attributable to the study, or to differences in baseline achievement levels, a highly significant factor.

Summary

- Students in pilot schools had significantly higher reading outcomes in 1999-2000 than students in control schools for:
  - 3rd grade (English version only) total reading scale scores, and
  - 4th grade (English version only) total reading scale scores.
- Students in pilot schools had significantly higher reading outcomes in 2000-2001 than students in control schools for:
  - 3rd grade (English version only) total reading scale scores;
  - 4th grade (English version only) total reading scale scores and Standards 1, 4, and 6;
  - 4th grade (English version only) total writing scale scores; and
  - 5th grade (English version only) total reading and math scale scores.
- Approximately one quarter of the variance in third grade pilot students' reading scale scores (26%) is attributable to classroom level differences, leaving 74% to be explained by child level differences;
- Sixty-four percent of the variance in fourth grade pilot students' reading scale scores is attributable to child differences, leaving 36% to be explained by classroom level differences;
- Third grade students in Approach One classrooms had reading scale scores that were 16.40 points lower than students in classrooms of Approach Three teachers and 9.98 points lower than students in Approach Two classrooms;
- Up to 61% of the original 36% of variance in fourth grade pilot students' reading scale scores attributable to classroom differences is explained by the approach and the quality of the teacher's objectives;
- On average higher reading scale scores were found:
  - For female students in pilot classrooms;
  - For students of higher English proficiency in pilot classrooms;
  - For children in the classrooms of pilot teachers who have been in the district longer;
  - For fourth grade students in the classrooms in Approach Two and Three schools; and
  - For fourth grade students in the classrooms of pilot teachers who received higher ratings for the quality of their objectives.
- On average lower reading scale scores were found:
  - For pilot Hispanic children when compared to other students in their classrooms; and
  - For pilot children of lower socioeconomic status (free/reduced lunch) when compared to children who were of higher socioeconomic status (pay for lunch).
- Together, child gender, English proficiency, ethnicity, and SES explain 10% of the original 74% of variation in third grade students' reading scale scores and 15% of the original 64% of variation in fourth grade students' reading scale scores attributed to child differences.

C. The 6+1 Trait Writing Assessment

The 6+1 Trait Writing assessment (Six-Trait) was developed in the 1980s as both a way to teach writing and a tool to assess writing. It is the foundation for the Northwest Regional Educational Laboratory's (NWREL) writing assessment model and the basis for the descriptive criteria used to define the qualities of good writing at different levels of achievement. Once the teachers know the traits well and have good consistency between rates and among groups, the link to instruction becomes clear. The six traits most commonly used are: ideas, organization, voice, word choice, sentence fluency, and conventions. An additional trait, presentation, is also part of the NWREL model but is not included in the program used by the Denver Public Schools.

The six traits used by the Denver teachers are:

- Ideas—the heart of the message, the content of the piece, the main theme, together with all the details that enrich and develop that theme.
- Organization—the internal structure of a piece of writing, the thread of central meaning, the pattern, so long as it fits the central idea.
- Voice—the writer coming through the words, the sense that a real person is speaking and cares about the message.
- Word Choice—the use of rich, colorful, precise language that communicates not just in a functional way, but in a way that moves and enlightens the reader.
- Sentence Fluency—the rhythm and flow of the language, the sound of word patterns, the way in which the writing plays to the ear, not just to the eye.
- Conventions—the mechanical correctness of the piece—spelling, grammar, and usage, paragraphing (indenting at the appropriate spots), use of capitals and punctuation.

The Denver Public Schools uses the Six-Trait as the district-wide test to assess students' writing skills and employs the Six-Trait analytic rubric in classroom instruction to score the assessment. The same rubric is used for all grades, adjusted to be age appropriate by the teacher. The NWREL website and materials provide samples of work at all levels to help teachers gain proficiency at scoring. Each of the traits is scored on a five-point scale:

1. Not Yet—a bare beginning; writer not yet showing any control.
2. Emerging—need for revision outweighs strengths; isolated moments hint at what the writer has in mind.
3. Developing—strengths and need for revision are about equal; about half-way home.
4. Competent—on balance, the strengths outweigh the weaknesses; a small amount of revision is needed.
5. Strong—shows control and skill in this trait; many strengths present.

Data Collection

For the purpose of this study, students' Six-Trait scores from the Spring 2000 and Spring 2001 administrations were analyzed. Files were merged and data for students who had scores at both administrations and who were in consecutive grades from 2000 to 2001 were included in the analyses. A total of 8,457 students were included in the analyses across grades three through eight. Three of the 12 elementary schools did not have two years of scores on their students. (Colfax did not have scores for the Spring 2001 administration. Columbian and Traylor had only scores for fifth graders in 1999-2000 and they did not have scores for 2001.) A total of 1,836 students were included in the pilot school analyses. Teacher information was also merged with the student data where available at the elementary level.

A variety of analyses were conducted using a change score—the difference between the students' score in Spring 2001 and Spring 2000. In order to assess the degree of change across a distribution ranging from –4 to +4, independent samples chi square tests were used to compare various groups at each grade level. The following hypotheses were tested:

- The distribution of change scores across the range from –4 to +4 was the same for both pilot and control schools.
- The students in the three approach groups are the same in the way they are distributed across the range of change scores.
- The distribution of change scores across the range from –4 to +4 for pilot school students was the same for teachers who included Six-Trait Writing as one or both of their objectives and those who did not include the Six-Trait Writing as one of their objectives.
- The distribution of change scores is the same for all 13 pilot schools.

While the analyses were conducted on the actual distribution of change scores from –4 to +4, the following comparisons present the percentage of students in each group scoring one or more points below the previous year, scoring the same as the previous year, or scoring one or more points above the previous year.

The number of students in the different groups varies considerably. For example, the range is 200 to 500 students in grades three to five in the pilot schools, and 500 to 1,126 in the same grades in the control schools. When separated by approach, some numbers become quite small. There are 14 third graders in Approach One schools and 141 fourth graders. The analyses below are based on the *percentages* of students in the various performance categories and not the *numbers* of students in each category.

Comparison of Pilot versus Control Schools by Grade

Figure 6-13 presents the performance of third through fifth grade students in the initial 12 elementary schools in the Pay for Performance Pilot compared to the students in the same grades at the initial 36 control schools. It also compares the performance of the sixth, seventh and eighth grade students at Horace Mann Middle School, the initial pilot middle school to the same grades at six middle control schools: Hill, Kepner, Lake, Merrill, Rishel and Martin Luther King, Jr. Numbers presented in bold represent the best results in each category (more students increased or fewer students decreased). Only grades where statistically significant differences were found are included in Figure 6-13.

Comparison of Three Approaches by Grade

Figure 6-14 presents Six-Trait results among the pilot schools by approach. There were four schools in each of the three Pay for Performance approaches, but Spring 2000 and Spring 2001 scores were not available for all of the 12 elementary schools; only Approach Three is comprised of four schools. Three schools were not included in these analyses due to lack of data. Colfax (Approach One), Columbian (Approach Two) and Traylor (Approach Two). In addition, the three Approach One schools included in the analysis had scores for only 14 third grade students. Results for the third grade are provided here, whenever the chi-square was significant, because the other two approaches had scores for a sufficient number of students to be meaningful. As in the previous figure, the bold indicates the greatest positive change (greatest increase or least decrease) for each grade.

Comparison of Use of Six-Trait Assessment in Teacher Objectives by Trait

Some teachers used Six-Trait in their objectives, while many did not. The comparisons in Figure 6-15 involve only teachers at the pilot schools in grades three, four and five for whom data were available, and compare the results for teachers who used Six-Trait in setting their objectives to teachers who did not. Teacher identification information was available for 2001 for a total of 41 teachers at these schools, in which 20 teachers included the Six-Trait Writing Assessment as the assessment in one or both of their objectives and 21 teachers who used some other measure (e.g., ITBS, grade Level Math, QRI, etc.). Students' scores were compared based on the number of points their scores increased or decreased or remained the same from the Spring 2000 administration of the Six-Trait to the Spring 2001 administration. A total of 276 students were included in the former group ("Six-Trait" group) and 768 students were in the latter group ("non Six-Trait"). Results are

FIG. 6-13

Comparison of Pilot and Control Schools by Grade and Trait

		Percentage Increase/Decrease	
Trait / Grade		Pilot Schools	Control Schools
Idea			
Grade 3	Decrease	**40**	48
	No Change	34	25
	Increase	26	**28**
Grade 4	Decrease	32	**26**
	No Change	35	31
	Increase	33	**44**
Grade 5	Decrease	27	**23**
	No Change	29	34
	Increase	**44**	43
Grade 8	Decrease	37	**25**
	No Change	41	32
	Increase	**34**	31
Organization			
Grade 3	Decrease	**34**	44
	No Change	36	26
	Increase	30	30
Grade 4	Decrease	31	**24**
	No Change	34	32
	Increase	35	**44**
Grade 5	Decrease	**23**	24
	No Change	34	33
	Increase	43	43
Grade 7	Decrease	**10**	28
	No Change	33	34
	Increase	**58**	37
Grade 8	Decrease	**27**	39
	No Change	37	31
	Increase	**36**	30
Voice			
Grade 4	Decrease	26	26
	No Change	37	32
	Increase	37	**42**

FIG. 6-13 CONTINUED

Comparison of Pilot and Control Schools by Grade and Trait

		Percentage Increase/Decrease	
Trait / Grade		Pilot Schools	Control Schools
Word Choice			
Grade 4	Decrease	25	**22**
	No Change	37	32
	Increase	37	**42**
Grade 6	Decrease	**24**	38
	No Change	47	33
	Increase	**30**	29
Grade 7	Decrease	**11**	27
	No Change	41	41
	Increase	**49**	32
Sentence Fluency			
Grade 4	Decrease	30	**24**
	No Change	37	35
	Increase	33	**41**
Grade 5	Decrease	24	**22**
	No Change	30	39
	Increase	**46**	39
Grade 7	Decrease	**15**	32
	No Change	35	35
	Increase	**49**	34
Conventions			
Grade 4	Decrease	36	**26**
	No Change	36	27
	Increase	28	**38**
Grade 7	Decrease	**11**	21
	No Change	34	37
	Increase	**54**	42

Bold figures represent the best performance: greater increase or less decrease presented by grade for those grades where a significant difference was found between the two teacher groups. Student data were not available for Colfax, Columbian or Traylor.

The following observations can be made concerning the use of Six-Trait for objective-setting (Figure 6-15). In general, teachers who used Six-Trait for one of their objectives had a larger percentage of students with increases and a smaller percentage of students with decreases than teachers who did not use Six-Trait. In this comparison the grade three "Six-Trait" group generally outperformed the "non Six-Trait" group across the six traits and had significant chi square differences.

Grades not shown in Figure 6-15 rarely had differences that were large enough to achieve significance, but the results of the Six-Trait group are being presented here because of the direct relationship to the Pay for Performance Pilot. The teachers in this group were specifically targeting student performance on this assessment as one or both of their objectives.

- In the case of the **Organization** trait, fifth grade teachers who included Six-Trait as one or more of their objectives had only 20% of their students with a decrease in performance. Thirty-five percent of the students had no change while 45% had increases of one or more points. The non Six-Trait group had a slightly higher percentage of students with decreases and lower percentages of students in the no change and increase categories.
- Significant differences were not found for grades four and five in any of the other traits. In the area of **Voice,** the fourth and fifth grade teachers in the Six-Trait group had 27% of their students scoring below their 2000 scores while 30% of the fourth grade and 40% of the fifth grade students scored above their previous scores.
- Fourth grade teachers did not fare as well on the **Word Choice** trait. Thirty percent showed a decrease compared to 20% of the fifth graders. A similar percent of fourth graders scored higher on the 2001 test while 47% of the fifth graders scored at least one point higher on the 2001 test.

FIG. 6-14

Comparison of Pilot Approaches by Grade and Trait

		Percentage Increase/Decrease		
Trait / Grade		Approach One	Approach Two	Approach Three
Idea				
Grade 4	Decrease	31	39	**26**
	No Change	33	31	40
	Increase	**37**	29	34
Organization				
Grade 4	Decrease	31	41	**20**
	No Change	32	33	37
	Increase	37	26	**42**
Grade 5	Decrease	**20**	29	21
	No Change	33	32	35
	Increase	**46**	38	44
Voice				
Grade 3	Decrease	50	**23**	44
	No Change	14	36	36
	Increase	36	**41**	20
Grade 4	Decrease	28	37	**15**
	No Change	34	33	44
	Increase	38	30	**41**
Grade 5	Decrease	**18**	36	21
	No Change	34	21	39
	Increase	**48**	43	39
Word Choice				
Grade 3	Decrease	**21**	23	39
	No Change	37	36	42
	Increase	**43**	41	19
Grade 4	Decrease	26	32	**17**
	No Change	35	43	38
	Increase	38	24	**45**

FIG. 6-14 CONTINUED

Comparison of Pilot Approaches by Grade and Trait

		Percentage Increase/Decrease		
Trait / Grade		Approach One	Approach Two	Approach Three
Sentence Fluency				
Grade 3	Decrease	29	**21**	42
	No Change	21	40	43
	Increase	**50**	39	16
Grade 4	Decrease	35	36	**20**
	No Change	29	40	40
	Increase	36	25	**40**
Grade 5	Decrease	**14**	34	23
	No Change	33	26	30
	Increase	**52**	39	47
Conventions				
Grade 3	Decrease	**21**	29	59
	No Change	29	30	29
	Increase	**50**	41	12
Grade 4	Decrease	38	39	**31**
	No Change	33	37	37
	Increase	29	24	**32**
Grade 5	Decrease	**16**	38	23
	No Change	39	26	42
	Increase	**45**	37	35

Bold figures represent the best performance: greater increase or less decrease

- About one third of the students in the classes of fourth grade teachers in the Six-Trait group had decreases in their scores on the **Sentence Fluency** trait while a quarter of the students scored above their previous year's score. Fifth grade teachers had only 19% of their students perform below their 2000 score while 48% had increases.

- On the final trait, **Conventions,** 39% of the fourth grade students had a lower score in 2001 and 27% had a higher score. The fifth grade teachers had 30% of their students with lower scores and 29% with higher scores.

Comparison of Six-Trait Results by School

Figure 6-16 presents the final comparison of the Six-Trait results, arranged by trait and by school. Because of the small sample sizes when breaking these down to the school level, all grades have been combined to produce a single school result. Figure 6-16 presents the percent increase, percent decrease and the percent of students with no change in score from 2000 to 2001 for the pilot schools for which data was available. The nine elementary schools comprise students in grades three to five and Horace Mann comprises students in grades six to eight.

Summary

The Six-Trait analysis yields a mixture of results. No clear conclusion can be identified between pilot and control schools overall, based on the results portrayed above. In some grades and for some traits, pilot school averages have improved more. For other grades and traits, control schools seem to have shown a greater increase. The same holds true for the differences between the pilot approaches.

An interesting finding is that scores increased for students whose teachers set one or more objective based on Six-Trait at about the same rate as students whose teachers did not include Six-Trait in one of their objectives. This result could be interpreted to indicate that using Six-Trait in teachers' objectives was not effective. However, it also indicates that a concern about teachers tending to grade students higher if they have bonuses tied to those grades is not borne out. In fact, the amount of growth measured by teachers is about the same, regardless of the focus on their objectives.

Assessments and programs like Six-Trait are valuable teaching tools, and can also be valuable in assessing student progress. The difficulties of using Six-Trait and similar assessments are discussed in Chapter IX. The extra effort it takes to score student essays reliably and fairly must be considered in using such an assessment for broad comparative purposes. On a limited basis, however, the use of such assessments helps balance an academic program, and steps should be taken to set up a reliable scoring method for this purpose.

FIG. 6-15

Comparison of Six-Trait Results of Teachers Who Included Six-Trait in One or More Objective and Teachers Who Did Not

		Percentage Increase/Decrease	
Trait / Grade		**Six-Trait as objective**	**Non Six-Trait objective**
Idea			
Grade 3	Decrease	**18**	51
	No Change	36	31
	Increase	**46**	18
Grade 4	Decrease	44	**23**
	No Change	30	39
	Increase	26	**38**
Grade 5	Decrease	30	**27**
	No Change	32	28
	Increase	39	**44**
Organization			
Grade 3	Decrease	**14**	43
	No Change	39	35
	Increase	**46**	22
Grade 4	Decrease	43	**22**
	No Change	31	36
	Increase	25	**41**
Voice			
Grade 3	Decrease	**18**	39
	No Change	39	34
	Increase	**43**	27
Word Choice			
Grade 3	Decrease	**9**	40
	No Change	43	37
	Increase	**48**	23
Sentence Fluency			
Grade 3	Decrease	**20**	38
	No Change	negligible	negligible
	Increase	**39**	22
Convention			
Grade 3	Decrease	**25**	53
	No Change	32	30
	Increase	**43**	17

Bold figures represent the best performance: greater increase or less decrease

FIG. 6-16

Comparison of Six-Trait Results by School

School		Idea			Organization			Voice		
	N	**-1**	**0**	**+1**	**-1**	**0**	**+1**	**-1**	**0**	**+1**
Cory	136	25	33	42	24	32	44	22	47	30
Centennial	222	32	30	37	29	32	38	23	33	44
Edison	127	49	33	18	54	28	17	53	23	24
Ellis	193	35	34	31	28	38	35	26	34	40
Fairview	128	30	30	40	22	39	39	32	31	38
Mitchell	153	34	29	37	29	35	37	24	41	35
Oakland	195	30	33	37	27	35	38	26	35	40
Smith	99	24	32	43	23	32	45	21	32	47
Southmoor	46	17	54	28	7	51	42	24	42	33
Horace Mann	450	34	34	32	26	33	40	30	36	34

School		Word Choice			Sentence Fluency			Conventions		
	N	**-1**	**0**	**+1**	**-1**	**0**	**+1**	**-1**	**0**	**+1**
Cory	136	26	37	38	26	36	38	26	38	37
Centennial	222	25	33	42	28	33	39	33	32	35
Edison	127	39	39	22	53	27	20	50	33	17
Ellis	193	28	37	35	31	40	29	43	37	21
Fairview	128	22	39	39	17	46	37	28	28	44
Mitchell	153	20	41	39	27	31	42	37	34	29
Oakland	195	21	34	45	30	31	39	34	37	29
Smith	99	21	32	47	17	31	53	13	36	51
Southmoor	46	9	24	67	7	38	56	13	50	37
Horace Mann	450	20	42	38	23	37	40	21	38	41

Percent of students scoring the same, one or more points above, or one or more points below their previous score.

CHAPTER VII

School and District Voices

A. Overview

Pay for Performance is rooted in the belief that administrators, teachers and students can form a learning partnership to improve student achievement. In this context, understanding school practice through school and district voices is an essential element of the pilot. It also provides insights as to the concerns and hopes of educators focused on improving student achievement.

This chapter summarizes the responses of teachers, administrators and parents at the pilot school sites with appropriate comparisons made to control schools. It also provides additional commentary from central administrators, board members, and other external community members.

B. A Methodological Roadmap

Interviews were conducted and surveys distributed to key constituency groups in each of the pilot's initial two years. The scope of survey distribution and the related response rates are delineated in Chapter III. These data were used to establish baselines. The second year data were also used to monitor changes in perceptions and expectations over the course of the pilot.

The first year survey (1999-2000) consisted of six sections. The first section asked respondents to provide background information, the second section dealt with participants' understanding of the goals of pay for performance, the third section covered participants' attitudes about teacher objectives, the fourth section dealt with participants' expectations for the pilot, and the fifth section dealt with the support participants believed would be necessary for the project to be successful. A sixth section, unlike the preceding five sections, consisted of a series of open-ended questions. These questions asked respondents to express, in their own words, what their goals and expectations were for the pilot.

The second year survey (2000-2001) contained the same sections as the first year as well as a section asking participants' attitudes about the implementation of Pay for Performance and a section asking participants to provide information about professional development activities at their respective school sites. Surveys were distributed to the initial set of pilot schools, new schools, incoming high schools, and a sample of control school teachers and administrators. A separate survey was designed and distributed to a random sample of parents in the pilot and control schools.

Face-to-face interviews were conducted with a range of community members, including teachers, administrators and parents at the school sites, central administrators, board members, association leaders and external community members in each of the two years. A primary goal of these interviews was to better understand the shared perceptions and contextual assumptions operating among and between constituent groups from year to year.

Whenever possible interviews were conducted one-on-one, though in some cases group interviews were conducted. Specific interview protocols were utilized to assure that a common set of questions was asked of interviewees. Further, because the interviews were interactive, protocol questions were a starting point and new questions were introduced during the course of the interviews.

Surveys were anonymous and interviews conducted confidentially. Survey information did include information necessary for the tabulation of survey results, such as school, position, grade level, length of service in district and at that particular school. Survey results were analyzed both quantitatively through ANOVA and chi square analyses for significant findings, and qualitatively utilizing coding categories to determine predominant themes and perceptions.

The use of a large number of variables yields large amounts of information, but makes analysis complex. For each of the 54 possible questions on the 2001 survey alone, there are at least 37 different combinations of responses. This number multiplies substantially when a second year's survey data are added. While all these data combinations are interesting, many will not enhance the main findings presented in this report. The tables and figures contained within represent data that illuminates general findings and offers contextual information necessary for interpretation.

The following comparisons were made:

- Among schools and approaches: 13 schools, three approaches to PFP;
- Among respondent groups: classroom teacher, special subject teacher, special education teacher, and school administrator;
- Among teachers who have varying years of experience: 1, 2, 3, 4-13, 14+;
- Between surveys conducted during the first year (1999-2000) and second year (2000-2001) of the pilot;
- Between pilot and control schools, where applicable;
- Between survey results and written or interview commentary;
- Between survey results and other observations, such as board or central administrator comments.

The Study Group

Before attempting to present findings, it is important to understand the community under study. Figures 7-1a-e delineate respondents for each of the two years' surveys according to the range of demographic and identifying data requested.

The majority of respondents in both years are classroom teachers. The largest percentage of teachers have been either in the school or district between 4 and 13 years and are DCTA members.

C. Perceptions of Goals and Priorities

The survey was intended to monitor changes between years and was divided into sections focusing on pilot goals, pilot support, teacher objectives, implementation, professional development, and impact. Change can be a highly personal experience and interviews and surveys provide the snapshot through which these changes can be monitored and understood.

At the school level, teachers and administrators were asked to identify the overall goals of PFP by agreeing or disagreeing with a list of possibilities gathered from conversations with the broader school community during early phases of the pilot.

Perception of Project Goals

Figure 7-2 summarizes the aggregate responses of teachers and administrators for each year. There are few significant changes in participants' perceptions of project goals between years. While most respondents indicate that increasing student achievement is an important goal of PFP, a majority also cite several other goals.

The greatest change is a drop in the percentage of respondents who believe that a goal of pay for performance is to change the system for teacher compensation, which drops from 80% of those agreeing in 1999-2000 to 68% agreeing in 2000-2001. This change suggests that respondents are changing their understanding of the intent of pay for performance.

Respondents also were less inclined to agree that pay for performance gave teachers and administrators tools to improve instruction (47%) in 1999-2000 and this perception dropped significantly in 2000-2001. Similarly, respondents rejected the view that a goal of pay for performance was to give administrators more control over teachers. While it appears as though respondents are clear about project goals, it should be noted that the joint goal statement for the pilot was not developed and adopted until early in 2001.

FIG. 7-1A

Pilot School Surveys by Respondent Group

Position	1999-2000		2000-2001	
	Number	Percent	Number	Percent
Classroom Teacher	190	63	232	64
Special Subject Teacher	11	4	47	13
Special Education	35	12	34	9
Specialist	28	9	32	9
School Administrator	18	6	8	2
Other	20	6	7	2
Total	302	100	360	99

FIG. 7-1B

Pilot School Surveys by Ethnicity

Ethnicity	1999-2000		2000-2001	
	Number	Percent	Number	Percent
Black	27	9	33	9
White	207	68	257	71
Hispanic	47	16	44	12
Asian	—	—	1	1
Multi-Racial	4	1	5	1
Other	12	4	13	4
No Answer/Multiple	5	2	9	3
Total	302	100	362	100

FIG. 7-1C

Number of Years in School and District, 1999-2000

Number of Years in the District	Number of Responses	Percent of Total Responses	Number of Years in the School	Number of Responses	Percent of Total Responses
One Year	30	10	One Year	61	20
Two Years	25	8	Two Years	30	10
Three Years	16	5	Three Years	31	10
4-13 Years	117	39	4-13 Years	127	42
14 or More Yrs.	95	32	14 or More Yrs.	28	9
No Answer	19	6	No Answer	25	8
Total	302	100	Total	302	99

*Longevity categories were established and are in use by the DCTA.

FIG. 7-1D

Number of Years in School and District, 2000-2001

Number of Years in the District	Number of Responses	Percent of Total Responses	Number of Years in the School	Number of Responses	Percent of Total Responses
One Year	52	19	One Year	90	25
Two Years	35	12	Two Years	53	15
Three Years	29	11	Three Years	33	9
4-13 Years	130	46	4-13 Years	153	42
14 or More Yrs.	32	11	14 or More Yrs.	32	9
No Answer	1	0	No Answer	1	0
Total	279	100	Total	362	100

Differences in Perception of Goals by Respondent Group

Figure 7-3 shows perceptions of pilot goals by respondent group. In this instance, aggregating responses revealed little variance between respondent groups. Generally, the responses from different site-level respondent groups cluster around the average with a range of five to eight percentage points on either side of the mean with a few exceptions. For example, the range of agreement as to whether focusing district activity on teaching and learning is a primary purpose of PFP is from 71% agreement for special education teachers to 87% for specialists. It is difficult to draw

FIG. 7-1E

DCTA Membership

Are you a member of DCTA	Number of Responses		Percent of Total Responses	
	Number 99–00	Percent 00–01	Number 99–00	Percent 00–01
Yes	185	61	227	63
No	96	32	134	37
No Answer	21	7	1	0

FIG. 7-2

Pilot Goals

Goal	Percent Agreeing 1999-2000	Percent Agreeing 2000-2001
Increase student achievement	84	82
Provide additional financial compensation to teachers	68	70
Focus district activity on improving teaching and learning	74	75
Provide additional motivation to teachers	73	74
Reward accomplishments in teaching	69	69
Increase teacher accountability for student achievement	78	74
Change the system for teacher compensation	80	68
Give teachers tools to improve instruction	47*	37
Give administrators tools to improve instruction	47*	38
Give administrators more control over teachers	39	41

*The 1999-2000 question asked whether the purpose was to give teachers *and* administrators a tool to improve instruction.

conclusions about the reasons for these differences since there appears to be general agreement among respondents with the differences only a matter of degree.

Qualitative survey comments would suggest otherwise. When teachers were asked to put into their own words their opinions and expectations for PFP, many indicated confusion about the role and intent of goal setting. As one teacher noted "I've seen teachers working toward their goals and there is more conversation about goals and lessons and activities to meet those goals," while another said "I still think it is useless. People are setting goals that are attainable for money with no regard to changing student achievement."

Others offered recommendations to improve the use of goals in the classroom by suggesting that "first and second semester goals would be more appropriate" or that more attention needs to be paid to monitoring goals. "Challenging goals were not written for all classrooms. Some teachers made low goals and received money while other teachers made high goals and didn't get the money. Monitor goals!"

Differences in Perception of Goals by School

School environments and goals vary and these differences can affect participants' perceptions in significant ways. Figure 7-4 explores differences in perception of pilot goals by school by the end of the second year. When compared with differences in perception by respondent group, a range of variation is noted. For example, while teachers at all schools tend to agree that increasing student achievement and providing additional motivation to teachers are PFP goals, there is considerably less agreement around whether providing financial compensation to teachers, increasing teacher accountability, or changing the teacher compensation system are goals.

Three schools, Cory, Oakland and Southmoor, express the majority opinion that giving teachers tools to improve instruction is a pilot goal. A significant number of other schools not only disagree with that goal, but also that a goal of PFP gives administrators tools to improve instruction or have more control over teachers.

Differences in Perception by School Approach

It has already been noted that the PFP approach chosen by each pilot school correlates significantly with school demographics, including performance on ITBS and CSAP. This is a significant finding because increases in student achievement at low performing schools may be more difficult to achieve and the stakes higher than in high performing schools. Demographics may impact differences in perception of approach by school. With this in mind, there is a difference in teacher perceptions by approach for both years as shown in Figure 7-5a, b.

Approach Two schools showed a significant increase in their understanding of the goals of

FIG. 7-3

Differences in Perception of Goals by Respondent Group

Goal	Classroom Teacher		School Administrator		Special Subject Teacher		Special Education Teacher		Specialist		Other	
Position	00	01	00	01	00	01	00	01	00	01	00	01
Increase student achievement	84	80	91	75	86	81	85	88	100	97	85	71
Change the system for teacher compensation	85	68	100	75	82	59	73	65	83	75	90	71
Focus district activity on improving teaching and learning	76	75	64	75	74	74	86	71	69	87	79	57
Provide additional motivation for teachers	74	74	73	100	71	76	68	65	67	77	85	71
Rewarding accomplishments in teaching	72	70	64	75	58	68	68	59	67	68	80	57
Increase teacher accountability for student achievement	78	76	73	63	88	64	82	68	53	75	85	86
Provide additional financial compensation to teachers	73	73	64	88	55	62	58	77	72	56	80	67
Give teachers and administrators tools to improve teaching	51	**	64	**	29	**	43	**	41	**	60	**
Give administrators more control over teachers	47	45	9	0	43	39	21	24	28	43	25	43
Give teachers tools to improve instruction	*	33	*	63	*	41	*	44	*	41	*	43
Give administrators tools to improve instruction	*	34	*	50	*	43	*	50	*	41	*	43

*Was not part of 2000 survey
**Was not part of 2001 survey

PFP between years while Approach One and Three showed decreases among most goals. All three schools decreased by almost half their perception that PFP gave teachers tools to improve instruction.

D. Teacher and Pilot Support

Figure 7-6 shows where teachers and administrators identified those areas in which they believed they needed support for the implementation of Pay for Performance in their schools and classrooms. In the 1999-2000 survey, nine areas of potential support and six in 2000-2001 were identified in four groupings—objectives, student achievement, instruction, and implementation.

In response to participants' recommendations, the Design Team and other Denver Public Schools departments (most notably Assessment and Testing) conducted a variety of training sessions. It is interesting to note that while the perceived need for support has decreased in each instance, approximately 60-70% of all respondents continue to indicate a need for support through training in all of the areas identified. As one teacher noted, "There needs to be more professional support for each teacher, especially new teachers, helping them to teach difficult and hard to reach children."

Teachers were also asked whether they had received professional development in areas that might help them to link objectives to student

FIG. 7-4

Perception of Goals by School, 2001-2001

School	Financial Compensation	Teacher Accountability	Compensation System	Teacher Tools	Administration Tool	Administration Control
Centennial (32)	87	78	83	38	38	45
Colfax (25)	68	60	79	28	32	48
Columbian (6)*	67	100	100	33	0	33
Cory (24)	67	88	78	54	46	54
Edison (34)	79	65	50	27	27	41
Ellis (28)	75	72	68	46	46	29
Fairview (22)	68	73	64	27	32	59
Mitchell (18)	100	61	72	17	22	44
Oakland (35)	94	83	64	57	60	22
Smith (30)	55	80	74	37	52	37
Southmoor (15)	80	73	40	60	53	33
Traylor (21)	64	82	67	33	23	32
Horace Mann Middle (20)*	35	65	69	25	30	55

* The number of respondents at Columbian and Horace Mann is significantly lower than the faculty/administrator populations at these schools.

learning targets and integrate content areas into objectives. These areas were rated by teachers as among the lowest with regard to professional development received. Interestingly, many administrators wanted support with classroom observation techniques, specifically in the area of linking objectives to school curriculum. Figure 7-7 summarizes professional development received.

Professional Development Received by School

Professional development is one area where information by school or across district was difficult to obtain. School site training, it appears, is highly idiosyncratic, sporadic and dependent upon individual school needs as determined by the principal, CDM, or faculty. Similarly, training provided by the Design Team and other DPS departments is voluntary at principal and/or teacher request. Understanding the content and impact of professional development and training on the writing of objectives, and using student achievement data for teaching and learning, will require further examination once the scope of training at particular schools can be documented. Before such analysis is undertaken, it is recommended that DPS conduct a professional development audit.

Figure 7-8 summarizes professional development received, and needed, for each pilot school. It should be noted that a majority of schools still perceive a need for professional development and/or support in writing objectives even though a majority already received support in that area.

Professional Development Needed by Teacher Longevity

Examining the perceived need for professional development of veteran teachers and those new to the profession may be useful particularly when implementing PFP at the school site. Figure 7-9 shows the relationship between the length of time they have been teaching in DPS and, specifically, how long they had taught at their particular schools to their perceived need for professional development.

With some exceptions, the need for training and support tends to decline as teacher experience increases. It is important to note, however, that more than half of even the most experienced teachers believe they need support in several areas

FIG. 7-5A

Goals by Approach, 1999-2000

Goals of the Pay for Performance Pilot	Approach		
	One	Two	Three
Increase in student achievement	88	73	91
Provide additional financial compensation for teachers	68	67	71
Focus district activity on improving teaching and learning	75	60	87
Provide additional motivation for teachers	73	67	80
Reward accomplishments in teaching	68	61	77
Increase teacher accountability for student achievement	81	67	84
Give teachers tools to improve instruction	81	82	80
Give administrators tools to improve instruction	50	34	57
Give administrators more control over teachers	41	41	36

to make PFP a success. Future analysis should examine teacher level of education with perceived need for support and/or professional development.

E. Objective Setting by Teachers

Pay for performance is predicated on the belief that teachers can write fair and reasonable objectives that are both attainable and measurable. They are the cornerstone of the pilot project as it is currently constructed. Teachers set objectives based on their knowledge of their students and curriculum and are subject to approval by the school principal. There are myriad difficulties monitoring the implementation of objectives. As previously described in Chapters III and IV, the adequacy of objectives thus far requires further examination and institutional response. Nonetheless, this system is designed to focus principals and teachers on the work of classroom instruction and the need for positive outcomes.

Teachers were asked a series of questions about objectives and objective writing in both years.

FIG. 7-5B

Goals by Approach, 2000-2001

Goals of the Pay for Performance Pilot	Approach		
	One	Two	Three
Increase in student achievement	83	85	80
Provide additional financial compensation for teachers	72	79	71
Focus district activity on improving teaching and learning	81	73	77
Provide additional motivation for teachers	78	81	69
Reward accomplishments in teaching	69	77	69
Increase teacher accountability for student achievement	77	73	73
Change the system for teacher compensation	71	68	67
Give teachers and administrators tools to improve instruction*	50	34	52
Give administrators more control over teachers	34	46	43

* Questions differed between years.

Figure 7-10 summarizes the responses of different professional categories or groups of respondents working at each school—classroom teachers, special subject teachers, special education teachers, specialists and school administrators. Upward changes of 10 or more percentage points from the first to the second year of the survey are highlighted in bold italic, downward changes of 10 or more percentage points are marked in bold.

- *Classroom teachers* show the greatest increases to questions about whether teachers can and will meet their objectives. They also show a smaller increase in confidence that teachers are setting challenging objectives. Teachers show a decrease in confidence in the fairness of objectives, although both of these areas still show the confidence of a majority of teachers.

- *Special subject teachers* show the greatest variation in response. For example, they show an increase in confidence that fair objectives can be set, but still rate this area the lowest of all groups. This is not a surprising response, as teachers of such subjects as music, physical education and art have received conflicting advice and instruction as to the basis for objectives—some have set their objectives based on ITBS improvements, whereas others have set theirs based on goals in their respective subjects. This area of need was identified much earlier by the Design Team and others, and steps have been taken within the district to create a uniform approach to objective-setting according to the content of the different disciplines. This process is far from complete, however, leaving special subject teachers with too little guidance and consistency. This issue may also be reflected in these teachers' responses to the question of classroom variations. The content on which to base objectives, more than classroom variations, is a major issue for this group.

FIG. 7-6

Areas of Support Needed

Area of Support	1999-2000	2000-2001
Objectives		
Greater clarity on how objectives should be set/measured	83	*
Ways to set objectives based on the needs of students	78	*
More training in objective setting	66	60
Student Achievement		
Greater access to student achievement data	69	61
Better understanding of student achievement data	69	61
Greater access to technology to analyze achievement data	71	*
Instruction		
Feedback on past success	83	*
Help in developing and implementing new teaching strategies	69	58
More time to analyze data and develop my skills	81	70
Implementation		
Support in managing a complex program	*	67

* Questions differed between years

FIG. 7-7

2001-2002 Professional Development

Area of Professional Development	Percent Who Received
Writing objectives	57
Using student achievement data	39
Linking objectives to student learning targets	23
Integrating content areas into objectives	27
Selecting and using appropriate strategies	34
Relating objectives to the school plan	34

- *Special education teachers* have expressed a range of concerns about PFP, depending at least in part on the individualized nature of their work. Some work with very low performing students, for whom the ability to spell their own name, for example, is a major achievement. Others work with many children who also participate in different mainstream classes. Like special subject teachers and specialists, it is not clear what the content of the objectives should be for many special education teachers. The relatively low responses as to whether teachers can set fair objectives, and whether all teachers can meet their objectives, are probably reflective of this problem.

- *Specialists* show a substantial drop in confidence that everyone can set fair objectives or that most teachers will meet their objectives. Specialists, who include nurses, psychologists and other non-teaching professionals, have had probably the greatest difficulty of any group in determining how to set objectives related to curricular goals, and while the Design Team recognized and initiated a district response to guide objective-setting for specialists, that process is far from complete. Thus, this reaction is a natural response to a process that has been difficult and confusing for this group. The large decrease in the need for training indicated by specialists is more surprising, but may indicate that the training already received has not met their need.

FIG. 7-8

Perceived Need for Professional Development/Support by School, 2000-2001

School	Writing Objectives		Selecting & Using Strategies		Use/ Understand Data	
	Rec'd	Need	Rec'd	Need	Rec'd	Need
Centennial (32)	63	41	28	63	59	58
Colfax (25)	56	60	40	64	56	72
Columbian (6)*	83	50	67	67	50	100
Cory (24)	67	58	54	44	50	50
Edison (34)	35	55	9	50	12	55
Ellis (28)	79	67	36	57	43	79
Fairview (22)	77	41	32	57	46	41
Horace Mann (20)*	52	63	10	47	10	58
Mitchell (18)	33	78	11	61	6	61
Oakland (35)	67	64	36	71	64	71
Smith (30)	55	65	42	61	36	50
Southmoor (15)	40	87	80	67	53	67
Traylor (21)	68	36	36	41	36	43

* The number of respondents at Columbian and Horace Mann is significantly lower than the faculty/administrator populations at these schools.

- *Administrators* appear to be less convinced in their own ability to ensure fair objectives, but show an increased confidence that objectives will nonetheless be fair. Their confidence that teachers can meet their objectives has grown substantially based on their experience to date. They also reflect a concern about issues of classroom variability, but this may be attributable to an initially unrealistic expectation rather than a dramatic increase in problems in this area.

Objectives by School and Approach

Figure 7-11 shows responses to questions about teacher objectives by school and by approach for the two years of the survey.

Given the variation across schools, specific reasons for the changes in opinion are most likely related to school-specific issues. Most questions show highly variable changes and none of the questions show an increase or decrease in teacher response exclusively across all schools. The following trends in responses are notable:

- *Professional development:* Teachers' responses to two questions about training to set and meet objectives support earlier indications that, despite ongoing professional development during the past year, teachers continue to feel a need for help in these areas. Several schools show a decreased need, which may indicate the effectiveness of past training or the extent to which they participated in that training, but

FIG. 7-9

Areas of Support Needed by Teacher Longevity, 2000-2001

	Years Teaching in DPS					
Area of Support	1	2	3	4–13	14+	Total
More training and support in objective setting	75	64	59	55	55	60
Greater access to student achievement data	67	66	72	58	58	61
Better understanding of student achievement data	79	69	68	54	57	61
Help in developing and implementing new teaching strategies	75	60	71	48	57	58
More time to analyze data and develop my skills	71	71	79	69	67	70
Support in managing a complex program	71	71	76	66	64	68

FIG. 7-10

Teacher Objectives by Position (Percent Agreeing)

Question	Classroom Teacher		Special Subject Teacher		Special Education Teacher		Specialist		School Administrator	
	00	01	00	01	00	01	00	01	00	01
Fair objectives can be set by all teachers	73	68	57	62	57	50	63	38	82	88
Principals ensure that teachers set equally challenging objectives	65	57	59	57	54	59	**74**	**55**	**91**	**50**
All teachers can meet objectives	***43***	***56***	50	59	43	44	50	56	***64***	***100***
Most teachers will meet their objectives	***71***	***82***	***69***	***89***	75	76	**86**	**71**	91	88
Teachers need training and support to *set* objectives	81	78	77	76	78	71	**92**	**81**	**73**	**63**
Teachers need training and support to *meet* objectives	76	69	77	73	74	76	**87**	**72**	**91**	**75**
Variations in classroom composition make fair objective-setting difficult	77	75	**83**	**73**	***68***	***85***	71	73	***40***	***88***
Most teachers use student achievement data to develop objectives	86	87	80	82	**96**	**82**	74	82	82	86
Most teachers set objectives that are challenging	74	78	***56***	***82***	***63***	***76***	73	68	***73***	***86***

the responses suggesting need are among the highest in this section.

- *Fair objectives:* While there is some decline, nearly two-thirds of the teachers continue to believe it is possible to set fair objectives. They are less clear that principals can or will ensure that objectives are fair.
- *Teachers meeting objectives:* 82% of respondents believe that most teachers will meet their objectives, up from 74% the first year. Two schools, Mitchell and Southmoor, show substantial drops in the number who believe teachers either can or will meet objectives. These drops are precipitate, and should be investigated. Both schools are Approach Three schools, but other Approach Three schools have shown increases in the numbers who say that teachers can and will meet their objectives.
- *Variations in classroom composition:* Teachers at most schools are concerned with this statement even more than in past years. This may reflect either their own experience trying to set objectives or a lack of trust regarding how achievement will be measured. Both lower and higher performing schools share this concern, which suggests that it is not an issue of comparing schools in different parts of town, but rather an issue at the classroom level.

- *Relationship of objectives to teacher evaluation and school planning:* The wide variation of responses indicates that different schools handle either these tasks, or the objective-setting process, differently. Many teachers are not involved and/or not aware of their school plans. In interviews, many teachers indicate no knowledge of their school plans; those that say their objectives are related tend to speak of a school-wide focus on a topic, such as writing or literacy. At least some schools have required one or both objectives in such a topic area, related to their school goals.

A few schools indicate a strong relationship between objective setting and teacher evaluation. Most indicate a fairly weak relationship.

Objectives by Longevity

As shown in Chapter IV and above, objective setting has proven more difficult and complex than many teachers initially believed. Figure 7-12 shows perceptions of teachers on objective writing delineated by years taught.

The differences among teachers do not correlate strongly with the number of years a teacher has been with the system. More than half indicate

FIG. 7-11

Teacher Objectives by School and Approach (Percent Agreeing)

Approach One Schools	Colfax		Oakland		Smith		Traylor	
Questions	00	01	00	01	00	01	00	01
Fair objectives can be set by all teachers	64	64	75	80	47	47	78	**62**
Principals ensure that teachers set equally challenging objectives	71	71	84	89	52	52	47	***57***
All teachers can meet objectives	44	***60***	55	60	32	**42**	36	41
Most teachers will meet their objectives	78	76	74	77	44	***68***	65	***86***
Teachers need training and support to *set* objectives	96	96	69	77	88	83	76	**55**
Teachers need training and support to *meet* objectives	71	**64**	75	74	87	87	68	**52**
Variations in classroom composition make fair objective-setting difficult	88	84	65	63	85	87	87	**71**
Most teachers use student achievement data to develop objectives	87	***100***	87	***97***	71	**83**	96	95
Most teachers set objectives that are challenging	75	75	69	***89***	56	60	78	***91***
Objectives and teacher evaluation are closely related	*	59	*	80	*	58	*	86
Objectives and the school plan are closely related	*	70	*	80	*	81	*	96

* Was not part of 2000 survey

FIG. 7-11 CONTINUED

Teacher Objectives by School and Approach (Percent Agreeing)

Approach Two Schools	Centennial		Columbian		Edison		Fairview	
Questions	00	01	00	01	00	01	00	01
Fair objectives can be set by all teachers	69	72	43	50	73	**62**	85	**68**
Principals ensure that teachers set equally challenging objectives	50	**39**	67	**33**	64	**44**	80	**52**
All teachers can meet objectives	22	***52***	38	33	41	***70***	30	***73***
Most teachers will meet their objectives	62	***93***	67	***83***	81	88	84	91
Teachers need training and support to *set* objectives	69	65	96	**83**	29	***63***	95	86
Teachers need training and support to *meet* objectives	66	74	86	83	46	46	90	81
Variations in classroom composition makes fair objective-setting difficult	86	87	73	100	68	68	70	77
Most teachers use student achievement data to develop objectives	90	90	96	100	86	94	85	86
Most teachers set objectives that are challenging	75	***87***	59	50	68	73	63	68
Objectives and teacher evaluation are closely related	*	43	*	100	*	46	*	68
Objectives and the school plan are closely related	*	41	*	100	*	55	*	82

* Was not part of 2000 survey

that all teachers can set fair objectives, and can meet those objectives. More than 80% at all levels of experience indicate that most teachers are using data to develop objectives, and while there is some variation the majority believe that the objectives set are challenging. As noted earlier, despite confidence in teachers' ability to set and meet objectives, large numbers indicate a need for training both in setting and meeting their objectives, suggesting that they believe there is still room for improvement.

F. Implementation

In the 2000-2001 survey, teachers were asked to respond to what they or their schools may be doing differently because of the pilot. These results may be taken to indicate both the impact of the pilot, and the extent to which being a part of the pilot has changed school and classroom practice. In the long term, if a school is not doing anything differently as a result of the pilot, PFP may not be considered a success. In the shorter term however, if schools in the pilot are not doing

FIG. 7-11 CONTINUED

Teacher Objectives by School and Approach (Percent Agreeing)

Approach Three Schools	Cory		Ellis		Horace Mann		Mitchell		Southmoor	
Questions	00	01	00	01	00	01	00	01	00	01
Fair objectives can be set by all teachers	86	**63**	86	**62**	51	47	87	78	82	**33**
Principals ensure that teachers set equally challenging objectives	75	73	79	**57**	59	**16**	53	56	64	60
All teachers can meet their objectives	52	***63***	59	***79***	41	37	80	***61***	73	***14***
Most teachers will meet objectives	86	91	83	***96***	67	***79***	90	**77**	91	**60**
Teachers need training and support to *set* objectives	90	**71**	79	**68**	91	**79**	87	89	91	93
Teachers need training and support to *meet* objectives	81	**64**	69	71	84	**68**	90	**79**	94	**80**
Variations in classroom composition makes fair objective-setting difficult	67	***79***	63	***75***	82	90	70	**56**	55	***73***
Most teachers use student achievement data to set objectives	86	**68**	93	85	78	83	83	78	80	87
Most teachers set objectives that are challenging	93	**83**	72	**93**	64	**53**	73	67	91	80
Objectives and teacher evaluation are closely related	*	48	*	59	*	47	*	56	*	67
Objectives and the school plan are closely related	*	75	*	68	*	42	*	53	*	100

* Was not part of 2000 survey

things differently as a result of being in the pilot, it may be because changes in school practice require consensus and individual classroom change is idiosyncratic. Another explanation could be that aspects of the pilot as it is currently structured may not be adequate to test its main concepts. This concern is discussed in Chapter VIII. Figure 7-13 shows the results that, in part, form this conclusion.

These results allow several different interpretations. First, the majority of respondents in most categories say both that they are doing what they have always done, and that they are not doing anything differently. At the same time, nearly half

FIG. 7-12

Objectives by Teacher Longevity, 2000-2001

	Years Teaching in DPS					
Questions on Objectives	**1**	**2**	**3**	**4–13**	**14+**	**Total**
Fair objectives can be set by all teacher	71	68	68	58	66	64
Principals ensure that teachers set equally challenging objectives	64	65	54	49	61	57
All teachers can meet objectives	60	70	54	52	58	57
Most teachers will meet their objectives	84	73	71	84	83	82
Teachers need training and support to *set* objectives	88	94	85	74	67	77
Teachers need training and support to *meet* objectives	86	94	85	74	67	77
Variations in classroom composition makes fair objective-setting difficult	77	66	75	69	61	70
Most teachers use student achievement data to develop objectives	80	84	85	83	89	85
Most teachers set objectives that are challenging	76	72	64	75	86	77

indicate that PFP has led to a greater focus on student achievement at their schools, a response that has been bolstered by many individual comments. Thus, at some schools at least, teachers perceive themselves to be preparing and teaching much as they have always done, but the context may be different. Preparing in a school where more data are available, for example, or where there is a greater focus on student achievement, may lead teachers to believe that the project has had an impact even though they believe they are not themselves doing anything differently.

Further, in the Expectations and Impact section of this chapter significant numbers of teachers indicate that the relationship of teacher and school practices to student achievement has improved,

FIG. 7-13

Implementation by Respondent Group, 2000-2001

Question	**Classroom Teacher**	**Special Subject Teacher**	**Special Education Teacher**	**Specialist**	**School Administrator**
There is a close relationship between objectives under PFP and teacher evaluation	62	54	47	62	75
There is a close relationship between objectives under PFP and my school's plan	70	62	63	69	63
PFP has led to a greater focus on student achievement at my school	48	47	42	46	63
I am doing things differently as a result of PFP	31	50	18	45	75
I am doing what I have always done to meet the objectives of PFP	79	69	74	70	50
PFP requires a lot of extra work	46	35	50	54	50

though this is not supported by the majority of interview and survey comments. Differences can and do occur when respondents are given the opportunity to voice their own opinions rather than in response to specific choices, such as survey questions. The suggestion is that some changes have taken place at the schools through the implementation of PFP beyond what was anticipated. Figure 7-14 shows responses to the questions by school.

With the exception of one school, Traylor, Approach One schools report that PFP has led to greater focus on school achievement, yet all schools report they aren't doing anything differently as a result of PFP and subsequently are doing what they have always done. This view is well documented in survey and interview comments. Approach Two schools report that PFP has not led to a greater focus on school achievement.

FIG. 7-14

Implementation by School and Approach, 2000-2001

Approach One Schools	Colfax	Oakland	Smith	Traylor
Objectives and teacher evaluation are closely related	59	80	58	86
Objectives and the school plan are closely related	70	80	81	96
PFP has led to greater focus on achievement	65	73	50	29
I am doing things differently as a result of PFP	39	38	42	14
I am doing what I have always done for PFP objectives	78	94	84	86
PFP requires a lot of extra work	50	29	52	33

Approach Two Schools	Centennial	Columbian	Edison	Fairview
Objectives and teacher evaluation are closely related	43	100	46	68
Objectives and the school plan are closely related	41	100	55	82
PFP has led to greater focus on achievement	29	33	47	41
I am doing things differently as a result of PFP	31	33	30	32
I am doing what I have always done for PFP objectives	77	50	73	82
PFP requires a lot of extra work	36	40	30	46

FIG. 7-14 CONTINUED

Implementation by School and Approach, 2000-2001

Approach Three Schools	Cory	Ellis	Horace Mann	Mitchell	Southmoor
Objectives and teacher evaluation are closely related	48	59	47	56	67
Objectives and the school plan are closely related	75	68	42	53	100
PFP has led to greater focus on achievement	38	61	47	35	87
I am doing things differently as a result of PFP	25	32	47	35	64
I am doing what I have always done for PFP objectives	78	67	65	65	57
PFP requires a lot of extra work	48	64	74	47	67

Most variation occurs in Approach Three schools with some schools indicating a greater focus on student achievement and some not. Only one school, Southmoor, reported doing things differently as a result of PFP. Unlike Approach One and Approach Two schools, several Approach Three schools said they thought PFP required a lot of extra work.

Implementation by Longevity

Few trends emerge from breaking down teacher responses by years in the district. Those teachers who have been teaching from four to thirteen years appear the most sure that they are doing as they have done in the past, and are the least convinced that PFP has led to a greater focus on student achievement. Figure 7-15 shows from one

FIG. 7-15

Implementation by Teacher Longevity, 2000-2001

	Years Teaching in DPS				
Implementation of PFP	**1**	**2**	**3**	**4–13**	**14+**
There is a close relationship between objectives under PFP and teacher evaluation	53	68	57	60	61
There is a close relationship between objectives under PFP and my school's plan	66	77	68	64	72
PFP has led to a greater focus on student achievement at my school	58	53	52	37	52
I am doing things differently as a result of PFP	52	35	33	28	34
I am doing what I have always done to meet the objectives of PFP	65	88	70	80	74
PFP requires a lot of extra work	46	52	48	43	46

half to two-thirds of respondents see a close relationship between objectives under PFP and both teacher evaluation and their school's plans. In general, it appears that the differences among teachers are more closely related to the nature of their position or the school they work in than to the number of years they have taught in Denver.

G. Expectations and Impact

Expectations

During the first year of the pilot, teachers and administrators were asked what they expected as a result of the pilot. Since all of the schools had voted to join, it might be reasonable to expect that the expectations would be more positive than negative, and in most instances they are. Whatever the reasons for their expectations, teachers at the pilot schools indicated those expectations during the pilot's first year, and had an opportunity to say what they thought had happened so far at the end of the second. Figures 7-16a-c show teachers' expectations (1999-2000) according to approach, school and position. Figures 7-17a-d display similar results for their perceptions of the impact of the pilot to date. These are shown by respondent group, longevity, approach and school.

There is majority agreement regarding pilot expectations in all schools from the three different approaches, with only one exception. When asked whether student performance will increase as a result of PFP, agreement varied by approach. Approach Two schools show the least confidence that this would occur although there was majority agreement that schools would focus more on student success. This suggests that the respondents do not expect a school's approach to have a direct impact on student achievement and that teachers are responsible. It also suggests that teachers may not be sufficiently prepared, or have the training and support necessary to link curriculum and objectives to classroom instruction that would result in increases in student achievement. A majority of respondents agree that it will be easier for teachers with high performing students to increase student achievement than under-performing students, a view that is supported

FIG. 7-16A

Expectations of Pay for Performance (Percent Agreeing), by Approach, 1999–2000

	Approach		
Expectations	**One**	**Two**	**Three**
Student performance will increase as a result of PFP	43	29	57
Most teachers will receive extra pay as a result of PFP	63	62	80
The school will focus more on student success as a result of PFP	50	43	59
Teachers will receive support needed to set appropriate objectives	53	49	68
Cooperation among teachers will be reduced as a result of PFP	32	45	24
It will be easier for teachers with high performing students to increase student achievement	60	70	50
It will be easier for teachers with underperforming students to increase student achievement	15	17	20
Teachers who have shown the least gain in the past will be allowed to set easier objectives	20	10	13
There will be no penalty for teachers who do not meet their objectives	46	47	48
Objectives based on standardized tests will force teachers to teach to the test	77	92	85
It will be difficult to set comparable objectives from one teacher to another	74	80	60
Teachers will have to modify their teaching style to meet their objectives	63	64	56
Higher teacher compensation will result in higher student achievement	33	25	33

FIG. 7-16B

Expectations of Pay for Performance by Position (Percent Agreeing), 1999–2000

Expectations	Classroom Teacher	School Administrator	Special Subject Teacher	Special Education Teacher	Specialist	Other
Student performance will increase as a result of PFP	44	60	47	43	29	53
Most teachers will receive extra pay as a result of PFP	68	73	77	68	81	75
The school will focus more on student success as a result of PFP	51	70	63	36	56	60
Teachers will receive the support needed to set appropriate objectives	56	91	56	59	65	67
Cooperation among teachers will be reduced as a result of PFP	36	36	32	36	28	42
It will be easier for teachers with high performing students to increase student achievement	64	64	61	70	63	40
It will be easier for teachers with underperforming students to increase student achievement	18	18	13	11	18	30
Teachers who have shown the least gain in the past will be allowed to set easier objectives	16	9	12	11	28	16
There will be no penalty for teachers who do not meet their objectives	48	73	44	39	56	53
Objectives based on standardized tests will force teachers to teach to the test	89	55	85	89	89	74
It will be difficult to set comparable objectives from one teacher to another	65	64	59	86	67	75
Teachers will have to modify their teaching style to meet their objectives	67	46	55	61	61	55
Higher teacher compensation will result in higher student achievement	32	27	32	29	22	25

in individual interviews and survey comments. A majority believe that cooperation between teachers will be reduced and there is little confidence that higher teacher compensation will result in higher student achievement.

An examination of expectations by position reveals little variation from that of approach, although in some instances administrators have different expectations related to the availability of support and the confidence that teachers will not experience penalties as a result of PFP. Teachers expect that PFP will increase the likelihood that they will teach to the test and most agree that they will have to modify their teaching styles to meet their objectives.

The results of this table were summarized by approach and are included so that individual schools can examine variations across schools.

Impact

In the second year of the pilot, survey questions were designed to explore the perceived impact the initial phase of the pilot produced in schools. The questions about impact are based in part on the original set of expectations, and in part on issues and concerns that arose during the initial phase of the pilot. Teachers were asked, for each item, to respond as to whether it had declined, stayed the same, or improved.

- *Student achievement:* Teacher responses to the questions are generally more positive than many had predicted. When these responses are examined by position, they follow the trends discussed earlier in this chapter. It is noteworthy, for example, that nearly half of all groups believe student achievement has increased in the past year, while only about 10% see a decline. While tests to date show little overall gain for pilot over control schools, the increased focus on achievement and additional support provided may have created an atmosphere more conducive to teaching and learning, even if student achievement results do not yet confirm this change. It is also significant that more teachers see an improvement in student achievement than expected it to happen.
- *Relationship of school practices to student achievement:* A third of all respondents see a greater relationship between school practices and student achievement, while only 4% see a decline. Classroom teachers see the least improvement in these areas, most likely reflecting past involvement in issues of achievement. It is significant, however, that 29% see improvement while only 3% see a decline. Similarly, all of the other respondent groups believe that there has been a substantial improvement in the relationship between school practices and student achievement. As with the previous question, approximately 50% of teachers said they expected a greater school focus on student achievement in 1999-2000, whereas a significantly higher percentage believe that this relationship has improved.
- *Relationship of teacher practices to student achievement:* A program that increases the focus on student success, when non-punitive, is more likely to demonstrate gains in achievement over time. In the context of PFP, the increase in the perceived relationship between teacher practices and student achievement is a positive, early development. It is interesting to note that the group that perceives the least improvement is classroom teachers, the group most focused on achievement in the past.
- *Relationship of professional development to student achievement:* These results mirror the questions above, confirming the expressed view that school focus on student achievement has increased.
- *Competition and cooperation:* The possibility that cooperation would be reduced and competition increased were among the greatest fears of teacher and administrators in interviews during the first year of the project. As the survey responses indicate, however, the large majority of all respondent groups indicate that there has been no reduction of competition as a result of PFP. As an unexpected benefit, a quarter of classroom teachers and a third of the specialists see improvements in cooperation among teachers, while most believe levels of cooperation have remained unchanged.
- *Support:* As with other questions, the majority of respondents report little or no change in areas of support but in most instances those who do see change see improvement. The most significant area has been in understanding student achievement although in all areas respondents indicate a need for increased, and increasingly focused, professional development.

Approach One Schools: When examined by position, teachers report little difference in perception with regard to student achievement. The majority of teachers in Approach One schools see an increase in student achievement with only 7% reporting a decline.

As previously reported, there continues to be a trend that while an improvement in student achievement may be perceived, there has been little change in school or teacher practice. A majority of teachers report that the relationship of school or teacher practices to student achievement has stayed the same over the course of the pilot.

FIG. 7-16C

Expectations of Pay for Performance by School (Percent Agreeing), 1999–2000

Expectations	Centennial	Colfax	Columbian	Cory	Edison
Student performance will increase as a result of PFP	26	42	45	67	33
Most teachers will receive extra pay as a result of PFP	68	75	67	71	67
The school will focus more on student success as a result of PFP	52	60	71	71	24
Teachers will receive the support needed to set appropriate objectives	39	61	71	80	50
Cooperation among teachers will be reduced as a result of PFP	43	21	50	14	41
It will be easier for teachers with high performing students to increase student achievement	82	70	76	33	57
It will be easier for teachers with underperforming students to increase student achievement	4	21	27	10	23
Teachers who have shown the least gain in the past will be allowed to set easier objectives	11	21	9	0	10
There will be no penalty for teachers who do not meet their objectives	54	24	50	43	50
Objectives based on standardized tests will force teachers to teach to the test	96	76	91	95	96
It will be difficult to set comparable objectives from one teacher to another	71	60	81	65	82
Teachers will have to modify their teaching style to meet their objectives	67	60	76	62	50
Higher teacher compensation will result in higher student achievement	35	20	29	33	23

There were slight differences in actual impact on cooperation and competition among teachers. While most teachers expected there to be increased competition, only 10% of teachers saw an increase while 23% saw an increase in cooperation. This was further supported by 23% of teachers reporting that communication between teacher and administrators had improved. A notable finding, which supports earlier findings, is that there is some improvement in support for PFP, improving classroom instruction, and understanding student achievement data. However, since the majority of respondents agree that it has largely stayed the same, the need for continued support is underscored.

Approach Two Schools: As previously noted in earlier findings, there is considerable variation across Approach Two schools on different variables. With regard to student achievement, two schools report a significant decline while at least one reports significant improvement. Similar variation exists with regard to the relationship of teacher and school

Ellis	Fairview	Mitchell	Oakland	Smith	Southmoor	Traylor
54	22	58	59	39	55	30
86	58	83	73	50	91	65
63	32	50	75	38	73	26
70	45	65	60	50	73	57
22	59	36	38	42	18	26
54	78	60	56	59	64	65
33	20	21	16	22	18	0
26	10	7	31	19	30	9
50	42	48	63	55	64	35
79	95	87	72	88	91	78
61	95	58	63	94	73	78
57	75	52	73	76	70	44
33	15	36	45	41	40	26

practice to student achievement with at least one school showing significant improvement in these areas and others showing decline or slight improvement. This pattern is noted throughout Figure 7-17c.

When compared to Approach One schools, there are several findings worth noting. While Approach One schools showed improvements in cooperation and declines in competition, Approach Two schools report the opposite. A majority of respondents report significant increases in competition among teachers and concurrent declines in cooperation. There is also a significant decline in communication. Similarly, Approach Two schools are experiencing a decline in support for PFP implementation, improving classroom instruction, and understanding student achievement data, with the exception of one school.

Approach Three Schools: Similar to Approach Two schools, it is hard to draw comparisons with so much variation between schools suggesting that

FIG. 7-17A

Teacher Perceptions of Impact by Respondent Group (Percent), 2000-2001

Question		Classroom Teacher			Special Subject Teacher			Special Education Teacher			Specialist			School Administrator		
		D	S	I	D	S	I	D	S	I	D	S	I	D	S	I
Student Achievement	D	10			14						9					
	S		44			35			53			46			29	
	I			47			51			47			46			71
Relationship of school practice to student achievement	D	3			10						5					
	S		68			54			61			48			57	
	I			29			37			39			48			43
Relationship of teacher practices to student achievement	D	2			10						10					
	S		67			44			61			45			57	
	I			32			49			39			45			43
Relationship of professional development to student achievement	D	5			10			7			5					
	S		70			64			61			65			71	
	I			25			27			32			30			29
Competition among teachers	D	6			13			7			14			14		
	S		86			75			81			76			86	
	I			9			13			13			10			
Cooperation among teachers	D	13			7			3			10					
	S		63			77			81			57			57	
	I			24			16			16			33			43
Communication between teachers and administrators	D	15			16			10			14			14		
	S		64			70			68			52			57	
	I			21			14			23			33			29
Support in implementing PFP	D	20			21			17			20			43		
	S		56			55			63			55			29	
	I			24			24			20			25			29
Support in improving classroom instruction	D	12			12			10			10			14		
	S		70			56			74			76			43	
	I			18			33			16			14			43
Support in understanding student achievement data	D	9			9			7			10					
	S		68			61			77			76			71	
	I			23			30			17			14			29

Key: D = Decline, S = Stay the Same, I = Improve

FIG. 7-17B

Teacher Perceptions of Impact for Approach One Schools (Percent), 2000-2001

Approach One Schools		Colfax			Oakland			Smith			Traylor		
Question		D	S	I	D	S	I	D	S	I	D	S	I
Student achievement	D				3			4					
	S		28			49			39			43	
	I			72			49			57			57
Relationship of school practice to student achievement	D							3					
	S		60			63			66			75	
	I			40			37			31			25
Relationship of teacher practices to student achievement	D							3					
	S		58			63			59			71	
	I			42			37			38			29
Relationship of professional development to student achievement	D				3			4					
	S		71			66			86			91	
	I			29			31			11			9
Competition among teachers	D	12			6			11					
	S		84			85			68			90	
	I			4			9			21			10
Cooperation among teachers	D	4			6			17			14		
	S		80			66			62			71	
	I			16			29			21			14
Communication between teachers and administrators	D							10			5		
	S		76			71			69			86	
	I			24			29			21			9
Support in implementing PFP	D	8			3			28			5		
	S		60			66			55			76	
	I			32			31			17			19
Support in improving classroom instruction	D	4						7			5		
	S		76			71			76			86	
	I			20			29			17			9
Support in understanding student achievement data	D	8						7			10		
	S		44			67			62			71	
	I			48			31			31			19

Key: D = Decline, S = Stay the Same, I = Improve

broader school and community influences may have an effect. Perceptions of respondents in Approach Three schools are more similar to Approach One schools, but for all schools there is more variance on the impact of PFP on student achievement and change in school and teacher practice.

Support has tended to decline in schools in all three approaches with individual schools experiencing differing degrees of support and/or professional development. This supports previous indications that professional development and support are idiosyncratic and subjected to in-school desire.

H. Comparison of Expectations to Perceptions of Change

The changing perceptions of the school community are an important lens through which to study pilot impact and its effect on school practice. The study monitored these changes through a series of qualitative survey questions in both years as well as through randomly selected interviews with teachers, administrators and parents at the school sites, central administrators, board members, association leaders, and external community members.

The questions asked in each year differed in several ways. First year questions were designed to establish a baseline of knowledge and perceived expectation about project goals. Second year questions focused on changes of opinion over time, level of agreement with the main PFP goal of linking compensation to student achievement, and recommendations.

In 2000, 171 teachers and administrators wrote in comments as did 151 in 2001. Interviews were conducted with 107 members of the Denver educational community in 1999-2000, and 213 members (including 43 parents) in 2000-2001.

Each of the surveys asks teachers and site administrators a variety of questions relative to their expectations, hopes and fears of the potential impact of the Pay for Performance Pilot. These are included not only in the Expectations and Impact section of the surveys, but were included in the interviews.

Student Achievement

Eighty-four percent of the teachers in 1999-2000 and 82% in 2000-2001 identify improving student achievement as one of the purposes of Pay for Performance. In 2000, teachers were asked to indicate whether they thought achievement would increase, and whether higher compensation would lead to such an increase. Overall, only about 40% of respondents indicated during the pilot's first year that they expected achievement to increase. Even fewer, about 30%, agreed with the expectation that higher teacher compensation would lead to greater achievement. As one teacher noted, "We are already working as hard as we could but why not get paid for something we were already doing?"

The line of thinking that many teachers follow is indicated in these quotes: to the extent that the purpose of Pay for Performance is a motivation for teachers to work harder, and therefore accomplish greater student achievement, they believe it is flawed. Thus, many have indicated that they do not believe PFP can result in higher achievement. Themes drawn from interviews and surveys support this view.

At the same time, a number of teachers in the pilot schools responded to the 2001 survey with the belief that student achievement had increased. In fact, in response to a question that asks whether student achievement has declined, stayed the same, or increased, 48% indicate a belief that it has improved, while 44% believe it has stayed the same. This might seem to contradict the initial belief reported in the survey comments and interviews. It may be that many teachers see PFP as more than a simple incentive—a concept which many still reject—and rather see a pilot that has had a variety of impacts, only one of which is compensation paid upon the achievement of certain objectives.

A third or more of respondents say that the relationship between school and teacher practices and student achievement has improved. Many indicate in interview and survey comments that they have achieved greater focus at their schools, at least in part a result of the pilot. Further, many teachers indicate that the conclusion that may be drawn is that, while teachers continue to harbor many reservations about the notion that incentive pay will improve student achievement, they may be saying that other activities related to PFP, from greater focus to a better understanding of student achievement data, have had a cumulative effect.

FIG. 7-17C

Teacher Perceptions of Impact for Approach Two Schools (Percent), 2000-2001

Approach Two Schools		Centennial			Columbian			Edison			Fairview		
Question		D	S	I	D	S	I	D	S	I	D	S	I
Student achievement	D	27									14		
	S		43			33			50			62	
	I			30			67			50			24
Relationship of school practice to student achievement	D	7									10		
	S		76			33			70			65	
	I			17			67			30			25
Relationship of teacher practices to student achievement	D	7									5		
	S		69			17			74			67	
	I			24			83			26			29
Relationship of professional development to student achievement	D	23									14		
	S		57			33			82			76	
	I			20			67			18			10
Competition among teachers	D	4						4			9		
	S		89			67			89			81	
	I			7			33			7			10
Cooperation among teachers	D	20			17			7			33		
	S		73			67			82			67	
	I			7			17			11			
Communication between teachers and administrators	D	47			17			18			48		
	S		47			50			68			52	
	I			7			33			14			
Support in implementing PFP	D	30			17			15			52		
	S		53			50			67			33	
	I			17			33			18			14
Support in improving classroom instruction	D	37			17			14			24		
	S		57						79			57	
	I			7			83			7			19
Support in understanding student achievement data	D	27						11			24		
	S		60			33			78			67	
	I			13			67			11			9

Key: D = Decline, S = Stay the Same, I = Improve

FIG. 7-17D

Teacher Perceptions of Impact for Approach Three Schools (Percent), 2000-2001

Approach Three Schools		Cory			Ellis			Horace Mann Middle			Mitchell			Southmoor		
Question		D	S	I	D	S	I	D	S	I	D	S	I	D	S	I
Student achievement	D				15			18			25					
	S		30			35			53			50			36	
	I			70			50			29			25			64
Relationship of school practice to student achievement	D				4			6			25					
	S		56			65			76			75			29	
	I			44			31			18						71
Relationship of teacher practices to student achievement	D				4						12					
	S		43			61			71			69			29	
	I			57			35			29			19			71
Relationship of professional development to student achievement	D				4			6			20					
	S		48			61			65			60			43	
	I			52			35			29			20			57
Competition among teachers	D	5			8						21			8		
	S		86			88			88			71			85	
	I			9			4			12			7			8
Cooperation among teachers	D							11			19					
	S		65			77			63			25			36	
	I			35			23			26			56			64
Communication between teachers and administrators	D				4			19			25					
	S		73			73			56			56			36	
	I			27			23			25			19			64
Support in implementing PFP	D	22			15			35			50					
	S		56			50			41			44			43	
	I			22			35			24			6			57
Support in improving classroom instruction	D	4			8			12			31					
	S		70			69			82			63				
	I			26			23			6			6		50	50
Support in understanding student achievement data	D				4			6			19					
	S		65			79			88			75			71	
	I			35			17			6			6			29

Key: D = Decline, S = Stay the Same, I = Improve

A predominant theme in the survey and interview data suggests an "I am doing what I have always done" belief. Teachers do not see PFP as significantly changing classroom practice. As one teacher noted: "These are the areas that we have been addressing for the last four years, not as a product of PFP, but as best practices for our community of learners." Many teachers talk about making an impact on students as the primary focus on teacher's work and that they do this "regardless of whether I have a written goal to accomplish the task." Administrators appear to see it differently. A majority of administrators believe that "PFP has brought more focus to issues, especially the need for goals, data awareness and better assessment tools." Despite the perception that the impact on the work of schools is generally small, many teachers and administrators suggest that the way PFP is structured may be the problem. One administrator said "PFP is not on center stage. It merits a big center stage. It should be the essence of how to run the district for kids," or as one teacher recommended, "Reward teachers and compensate them for hard work done. Too much emphasis is being put on performance. There are no rewards for the process. It is becoming an 'I will win' at all costs. It borderlines unethical teaching. Kids aren't a commodity."

It is interesting to compare these responses with Year One expectations. When teachers were asked why they voted in favor of PFP many felt they wanted to "explore new methods" and saw it as "a way to provide compensation for teachers who are working to improve student achievement." Many also expressed an interest in being part of the process and having the opportunity to give input into a system that might benefit teachers in the long run. Few teachers mentioned money as a reason for participation and this view has held consistent through comments expressed in Year Two.

Teacher Objectives

Writing fair and reasonable objectives is fundamental to the success of PFP as it is currently structured. And, as expected, teachers are concerned about a growing distrust in school surrounding this issue. Seventy-three percent of teachers believe that fair objectives could be set in 2000, but that confidence decreases to 68% in 2001. More teachers are confident in 2001 than were in 2000 that all teachers will meet those objectives. Teachers also show increases in requiring support to write fair objectives, and specialists and special education teachers continue to struggle with the adequacy of their own objective-writing. Teachers also express the least confidence that it would be easier for teachers with under-performing students to increase student achievement, a factor that is not so easily rectified under the PFP plan as it is currently structured.

Fairness is an issue that threads its way through survey and interview comments across all sectors in addition to survey data. While perceptions of fairness vary, clearly school community members are concerned with the ability to set fair and reasonable objectives that can be monitored objectively and assessed accurately. Several recurring concerns surfaced and included: the range of diversity of students in individual schools, classrooms, and across the district; the ability of teachers to set lower expectations and/or manipulate outcomes for compensation; the lack of standardized measurements related to objectives; teacher performance as measured by student performance; and the use of standardized tests as an acceptable measure of achievement for all students.

As one teacher noted, "I continue to feel that success in meeting objectives is a function of how a teacher writes their objectives rather than variables such as teaching methods or how hard a teacher works." Or, responding to diversity, one teacher said, "I thought it was a good idea at first, but there are so many variables that make it difficult to control outcome. Emphasis on scores versus other important goals such as parent/student satisfaction, student retention are lost."

Some of the fairness issues are real ones with potential solutions. It is possible to find ways to address issues of consistent measurement and the development of objectives appropriate to a particular position within a school, such as specialist or special education teacher. Those more difficult to address will continue to be difficult, such as the range of diversity within school with regard to objective setting.

Concomitant with issues of fairness is the resultant perception that unfair practice can engender a climate of distrust. And, not surprisingly, many respondents in the 2001 survey mentioned that the implementation of PFP is creating a climate of distrust and competitiveness in their schools. Many teachers suggested they now believed their colleagues would falsify records and/or cheat to meet objectives. When commenting on PFP as an incentive, one teacher said, "The only incentive it will provide will be to either: (1) set lower, reachable goals, (2) create animosity among teachers who won't want low achieving students in their room, (3) encouraging teaching to tests, which won't benefit students in the long run." Others documented unethical practices resulting from PFP and cited specific examples, "Some teachers at my school actually altered the achievement of their students to get the money. As my ethics kicked in I realized this is not right to compensate teachers for their job. We are here for the kids, not the money."

While much discussion about introducing competition into schools has proliferated over the years, there has been little evidence that it produces positive changes in teaching and instruction. Teachers regularly commented on this stating that, "If the faculty doesn't have common goals, the competition between teachers will result in a divided faculty." Many administrators concurred saying they were "worried about divisiveness as the project continues." Yet others, particularly at the school site level, continued to see the merits of PFP as a practice that kept "everybody on the same page" and saw distrust as the "real cost of major change." While issues of distrust can have consequences, they can be addressed through the careful monitoring of teacher objectives and their measurement, continued communication, and increased training and support during implementation.

Teacher Support

Teachers seemed to struggle with, "I'm doing what I've always done," and, "I need more support." Survey data from the pilot schools suggests that while training was received for writing objectives, selecting and using strategies and using and understanding student achievement data, more support is needed (refer to Figure 7-8). Support was defined in additional ways in survey and interview data and included requests for more information and training from the Design Team, an understanding of the "true" nature of teachers' work from school site and district administration, assistance developing "reasonable" objectives. Teachers and principals, in general, though admitting to an introductory session by the Design Team that was fairly helpful, seem to want more. Additionally, there seems to be consensus that Design Team trainings were more focused in the second year, though more training was needed. As one board member noted, "I haven't heard of any activities at the schools to help people to meet objectives. The board hasn't had the discussion. The teacher leaves the principal's office. What support is then given to these personnel? If there isn't support, it's 'business as usual.'"

While it appears as though a fair number of teachers suggest in 2001 that they "are doing what they've always done," many report an openness to participating in pilot goals in the 2000 survey and emphasize the need to examine student achievement. It appears as though the pilot, as implemented, did not meet their initial expectations particularly around the writing of objectives and the monitoring of those objectives. Yet some administrators look to teachers to work together toward the creation of common goals. One administrator said, "The teachers all have a common 40-minute planning time across grade levels. They can use that time to develop common goals for grade levels. Also, we have several teams and each team has a representative for different grade levels." Teachers also lacked training and support for the implementation of PFP goals, though it is hard to document the full extent of that lack of training since records of training and professional development activities are limited to self-reports. As one administrator recognized, "Developing common goals is tough; also developing meaningful goals. We need to decide which goals are reachable and which goals are realistic."

The implementation of PFP seems to require school-wide cooperation and support. Teachers mention the effects change in leadership has on the implementation of school-wide goals, or the emphasis on PFP in their schools. "Having served under two different principals, I can see a big difference on what each principal expects and considers necessary to actually receive money for making the goals," one teacher noted. Others said that it would be "more beneficial to students and teachers if our in-house administration were to remain more stable so our expectations remain consistent." This same concern is echoed concerning the district level where principals and teachers maintain that consistency is imperative. A teacher notes, "We need to be more informed about PFP. What do they want to do? All I know is that we write goals and then meet them."

Teachers initially were concerned that cooperation and competition would decline but these expectations were not achieved with the exception of selected schools. Overall through PFP, teachers have found ways to cooperate and competition has either stayed the same or decreased slightly. When asked to comment on ways in which PFP has changed relationships within the school, one teacher gave this example: "After achieving last year's goals, we as a team, because we do all of our planning together, decided it was too easy to get the money and we wanted to do something different. We've had the same goals so we made the goals harder for this year. We may have made one of them too hard, but that's OK. PFP has caused us to 'raise the bar'. "

Teacher Compensation

Sixty-eight percent of teachers in 2000 and 70% of teachers in 2001 agreed that PFP would provide financial compensation for teachers. Agreement as to whether or not PFP would change the compensation system for teachers dropped from 80% in the first year to 68% in the second. Similarly, when asked to rate whether or not higher teacher compensation will result in higher student achievement, teachers in all schools rated this as one of their lowest expectations.

Confusion reigns regarding what PFP is supposed to do by way of teacher compensation. Some teachers and administrators believe that DPS is trying to develop a new compensation system, yet others view it as a "bonus" for good teaching rather than an "incentive" to develop better teaching practice. Still others see it as punitive saying that "it [PFP] will only punish teachers with low achieving classes or children without help from home or special classes." The remaining group of teachers suggest that "teachers aren't in it for the money, their hearts are in it for the kids" and dismiss the idea of extra compensation altogether.

When teachers were specifically asked whether linking compensation to student achievement will provide additional incentive for teachers to increase student achievement, the majority said that teachers already work to improve student achievement and will continue to do so with or without compensation, especially when the compensation is so low. This suggests that either teachers aren't aware of how what they do in the classroom is linked to student learning, or that the compensation simply isn't enough to make a difference in altering classroom practice.

This level of uncertainty concerning the role of compensation also emerges when teachers and administrators are asked what they would include should they be designing an incentive system for teachers. The majority wanted a good salary, rather than a small "bonus" plan. Interestingly, many teachers mentioned things other than compensation like more planning time, more professional development, a fair and equitable way of measuring student growth for all students, and other non-financial systems of compensation. As one teacher noted, "In a perfect world, incentives would work. But in a perfect world, who would need incentives? The students would be self-motivated! Force the community and parents to be involved. That is the only way."

I. Parent Perceptions

A survey was sent to a random sample of 1,200 DPS parents from the pilot and control schools in English and Spanish. In addition to background information, parents were asked a series of questions about how they had learned about PFP, from whom, and whether or not PFP was taking place at their child's school. Parents were also asked to identify PFP goals and respond to a series of questions about what they expected to see as a result of PFP. Interviews were held with 43 parents, and parents were asked to comment on whether or not teacher compensation should be based on student achievement. Figures 7-18 to 7-20 present those responses.

The survey response rate was 12%, a much lower response than was anticipated. This could suggest that there is little knowledge about PFP

FIG. 7-18

Parents Responding by Grade Level, 2000-2001

Grades	Number of Parents Responding*
K-2	74
3-5	65
6-8	27
9-12	15

*The number responding by category exceeds the total number of surveys received due to parents having more than one child in the district.

FIG. 7-19

Goals of PFP, 2000-2001

Goals	Number of Parents Agreeing
Increase student achievement	88
Change how teachers get paid	71
Provide greater motivation for teachers	87
Increase teacher accountability For student achievement	95
Provide more pay for teachers	65

among parents and, therefore, little incentive to return the survey. This is supported by survey findings in which 40% of the parent respondents indicated they knew "nothing at all" about PFP. Sixty-seven percent of parents responding didn't know if it was taking place in any of the schools their children attend.

Parents were asked to comment on the overall goals for Pay for Performance. Clearly, of those parents responding, the majority believe that PFP is designed to increase student achievement, improve teacher accountability and provide greater motivation for teachers. One parent said, "I think teacher motivation will improve if their success with certain goals is required," or as another said "It's a really good thing to have accountability. Those who do well ought to be rewarded. I would like to think that good teaching is rewarded."

Parents are less inclined to see PFP as changing how teachers are paid, although 68% of parents believe that teacher compensation should be based in part on student achievement. But, as one parent cautioned, "Such a linkage must be done carefully so it doesn't simply reward teachers with gifted students or students with highly involved parents, or students from advantaged socioeconomic backgrounds."

It is difficult to have an expectation for a program you know little about, but parent expectations were fairly evenly distributed with the expectation that PFP would result in school change. Parents, like teachers, seem to believe that implementing PFP will assist schools to focus more on student success as a result of PFP.

J. Summary of Responses, Issues and Concerns

As the survey data and interviews would suggest, the implementation of PFP is a complex and dynamic process which, at this juncture, represents a point on a continuum of change. The results contained in this chapter indicate that, thus far, the impact of PFP is not startling, but change is evident and that change is largely positive. It is clear from the data that schools are indeed focusing more on student success.

Teachers want more support, particularly in the area of objective writing and using student data to inform classroom practice and also with developing strategies to reach those hardest-to-reach students.

While there was much concern over how PFP would impact school climate, those fears seem to be diminishing and are being replaced with examples of cooperation among teachers and collaboration with common goal setting. Teachers seem to want to improve practice and are looking for school and district support. To date, a comprehensive professional development and support system to complement PFP is not in place, at least in teachers' eyes. This type of system would enhance their ability to make increases in student achievement. Teachers are particularly concerned about the fairness of objective setting, and this is a reasonable concern given the wide range and scope of objectives. This, coupled with the need for multiple assessment measures, is an area that Design Team members and CTAC will examine in future years.

It is not clear from survey data that particular approaches have a greater or lesser impact on the implementation of PFP, rather broader student and institutional factors seem to have more effect. Further analysis during the second half of the pilot is required to determine the extent of impact these factors can and will have on increasing student achievement.

The school community offers a rich blend of feelings, desires and hopes for the future of DPS and, in particular, the role of PFP in their schools. Interviews and survey comments are laden with insight. In sum total, the experiences of those willing to be interviewed or responding to surveys offer veteran administrators and teachers, as well as those new to the district, concerns and ideas that can provide the roadmap necessary for continued improvement and success.

FIG. 7-20

Expectations for PFP, 2000-2001

Expectations	Number of Parents Agreeing
Student achievement will increase as a result of PFP	68
The school will focus more on student success as a result of PFP	62
Teachers will have to change their teaching styles to meet their objectives	68
Teachers will receive more pay	69
PFP will not result in much change at the school	36

CHAPTER VIII

Institutional Factors and Adjustments

A. Introduction

There are major institutional and external factors that have significantly effected the implementation of the pilot. Some continue to do so. The district and Design Team are seeking to address these concerns, although many are not unique to Denver. Instead, these challenges are largely endemic to urban school districts throughout the nation. This chapter describes many of the salient factors which have shaped, slowed, or enhanced the pilot's first two years of implementation. These have influenced the results of the pilot to date, and form the basis for district action in the coming years.

B. The Context for Experimentation

The Board of Education and the Association demonstrated courage and creativity in embarking on such a bold pilot. The many unique elements of the pilot are noted throughout this report. One of the earliest pilot learnings which emerged for the parties is that they, and the district as a whole, underestimated the complexity and scope of the experiment they were undertaking. Whereas a new dropout prevention program might be piloted on a specific audience of students at a single high school, Pay for Performance is an attempt to align achievement, instruction, assessment, professional development and compensation. Even in the pilot stage, it is a significant undertaking and requires a substantial district commitment.

At first glance, the concept of pay for performance appears to be simplicity itself. Indeed, several of the earliest proponents in Denver saw the pilot as a simple experiment. However, the sponsoring parties soon realized that the pilot faced distinct organizational challenges. For example, any effort to examine a teacher's impact on individual student achievement requires much greater data access and understanding than had previously been available to the sites, including data on many different types of assessments. Further, the challenges of alignment—as reflected in the need to link objectives to district standards to expectations for teachers to student results—are over-arching for a large district. Experimenting with these components at a relatively small group of schools still requires that the system be able to make and lead changes that ensure the quality and integrity of the pilot. This acknowledgement framed many of the sponsors' subsequent decisions and actions.

C. Leadership: Changes v. Stability

There is characteristically a direct relationship between the quality of an organization's leadership and the results which the organization is able to achieve. In this regard, some of the pilot's weaknesses are also in areas of strength. This is particularly apparent around issues of leadership.

Management

Since the pilot was formulated in 1999, the superintendency of the Denver Public Schools has changed five times. The first superintendent during this period was one of the primary thinkers behind the concept of pay for performance. However, he announced that he was leaving in February 1999 and subsequently retired in June. At this point, contract negotiations were underway. One of his assistant superintendents then served as acting superintendent until August. This covered the critical period of negotiating the terms of the pilot. The next superintendent was appointed in July and began work in mid-August 1999. His superintendency became embattled relatively quickly and the Board of Education made a change in leadership in May 2000. Then, the assistant superintendent of secondary education was made the fourth superintendent during the pilot period, initially in an acting capacity. She was named interim superintendent in June 2000. Finally, in April 2001, a new superintendent was chosen from a related but different field (community colleges) to be the district's fifth superintendent in this two year period. He assumed the position formally in June 2001.

Having five chief executive officers in a two year period creates strains in any large organization. In particular, during the initial two years of the pilot, there were two acting/interim superintendents and one permanent superintendent whose tenure was extremely short. It is not a criticism of any individual to note that this created a situation of weakened institutional leadership at a time when substantial change and significant executive leadership were needed. Such leadership turmoil at the executive level is increasingly common in urban school districts. It dilutes the efforts of any district to implement meaningful change. In Denver, this has constrained the best efforts of many individuals to institutionalize the pilot within the district. Several of the ensuing problems, and the need and efforts to overcome these problems, are discussed in the balance of this chapter.

Board of Education

Leadership from the Board of Education has been far more consistent than at the executive level of the district. Since the start of the pilot there have been just a few changes in board composition. This included the decision by the board's leading proponent of pay for performance not to run for re-election. In addition, there was a change of board officers. However, the board's commitment to pay for performance has not wavered.

During the initial two years of the pilot, the board has had to deal with a range of exigencies. These have ranged from the need to maintain the district's focus during the leadership turnover described above to the requirement to make necessary adjustments to ensure the viability of the pilot. Throughout this period, the board has been resolute in stating, and re-confirming when necessary, that pay for performance is one of the district's highest priorities. In fact, when hiring the most recent superintendent, the commitment to pay for performance was one of the key criteria used by the board when evaluating candidates.

Denver Classroom Teachers Association

The Association has also been steadfast in supporting the pilot. In part, this has resulted from a consistent leadership structure. The initial two appointees to the Design Team were, respectively, the Association's Vice President and the lead contract negotiator. The Vice President, in turn, became President of the Association at the end of the pilot's second year. This was a smooth, planned transition. The Executive Director of the Association has served in this position both before and during the pilot. Also, reflecting the Association's participatory structure, building representatives and Association leaders have been regularly briefed on the pilot.

This stability has been essential. Although not a replacement for consistent executive leadership, it has helped the pilot to advance despite the strains resulting from the turnover in the superintendency. Moreover, at various junctures when corrections or modifications were required to strengthen the pilot, the Association has demonstrated an unusually high level of collaboration with the Board of Education.

D. Purpose

During the first two years of the pilot, a need emerged for greater clarity regarding the purpose of the pilot. Interviews, survey responses and written records confirmed the importance of clarifying the purpose of the pilot. Members of the Board of Education, DCTA leadership and bargaining team, and other key figures within the administration have all described their understanding of the purpose of the pilot. These understandings were not formalized in the initial contract in a way that best captured the intent of the pilot. This subsequently became one basis for the Board of Education and the Association to amend the contract.

The initial 1999 contractual language states: "The Board of Education and the Association agree to collaboratively design a performance pay plan for teachers." Specifics agreed to during the negotiations, such as the use of teacher-set objectives, the three different approaches and the Design Team, are defined within the contract. However, the overall purpose is, at root, assumed. As a result, different individuals and different constituencies had varying understandings and, therefore, expectations of the pilot, from "I like PFP, it's a nice little benefit for what we are already doing," to "The bottom line is how does PFP drive improvement," to "We need to change the entitlement mentality in teacher compensation," to "Prove it. Prove PFP works."

The need to clarify the purpose was identified during the pilot's first year by both internal and external audiences. As noted in Chapter II, there were discussions on this issue from June through December 2000. In a focused effort to remedy this problem, a formal statement of purpose was agreed upon and then issued in January 2001. This was then incorporated into the contract.

The Design Team has since taken on the lead charge of communicating the purpose of the pilot to the district and community. This is a pivotal function. For example, as shown in Chapter VII, 80% of the pilot school teachers surveyed agreed in the first year that changing the system of teacher compensation was a primary pilot goal. In year two, though, only 68% see this as a primary goal. This is a change in perception that is worth exploring by the district.

This lack of clear purpose has hindered the implementation of the pilot. Several related issues are noted below:

- The structure of the pilot tests the ultimate purpose in a narrow construct. Whereas the statement of purpose speaks of "altering the salary structure" and using student achievement as "a part of the compensation system," the pilot itself is based on providing relatively small bonuses on top of the existing, traditional compensation system. The results which emerge from the pilot will be based on this construct. While much will be learned as a result of the pilot, and much of this learning will be applicable to a revised compensation system, critical aspects of the compensation system have not yet been developed and are therefore not being tested. The district will need to ensure that the Joint Task Force on Teacher Salary, a recent adjunct to the pilot, bridges this gap.
- Teachers and administrators participating in the pilot have had varying understandings as to the purpose of the pilot, as described above and in Chapter VII. Some, but not all, have understood that the structure of the pilot is

not necessarily the final structure of a pay for performance compensation system. Nonetheless, differing understandings as to the pilot's purpose have confused many teachers. This has led some to vote for or against participating based on speculation as to the pilot's intent.

- The lack of clear purpose has also affected administrative support structures. It has been noted repeatedly during the course of the pilot that functions and structures of the central administration have not been sufficiently aligned in support of the pilot. Part of the reason for this lack of alignment is the lack of clear purpose. During much of the two-year period covered by this report, many central administrators identified the Design Team as the managers of the pilot, indicating that the pilot was not the responsibility of their departments. "We operate as little islands and there is no boat to go between the islands," observed one administrator. Or as another administrator said, "PFP is important, but we have many other things of importance, particularly state mandates." It also contributed to a feeling among many that "this too shall pass." Another central administrator said, "There's this feeling that we'd just as soon it go away," and "It causes pain and we don't own it." Yet the work needed in central departments to fully support and align behind this pilot is substantial.

E. Pilot Extension and Baseline

Extension

The pilot was initially proposed and endorsed as a two-year initiative. It was negotiated in Spring 1999 and ratified in September 1999. The first year of implementation was to be 1999-2000. Thus, the pilot was underway before it was fully planned or designed.

As indicated in Chapter II, it soon became apparent that the district would need to develop greater institutional capacity in order to implement a pilot of quality. Further, additional pilot years would be needed to measure the impact of the pilot. Rather than react defensively, both the Board of Education and the Association examined the merits of extending the pilot. This resulted in the formal extension of the pilot to four years. This extension, in turn, has made it possible to begin to address problems of district infrastructure and to generate external philanthropic support in support of the pilot.

Baseline Year

Under these circumstances, it became clear that the first year of the pilot was, in reality, more of a developmental year than a full test of the concept of pay for performance. Particularly given the mutual interest of the Board of Education and the Association in benchmarking progress and conducting a comprehensive study of pilot impact, it was essential to establish a baseline for measurement purposes. This would require substantial, high quality baseline data. As a result, the 1999-2000 school year became the baseline year for pilot implementation and for the study.

This was an operational decision of the Design Team which the Board of Education and the Association supported. With such a baseline, it would be possible to examine pilot impact and frame conclusions regarding the factors contributing to that impact.

F. Problems of Implementation

The pilot has provided a vehicle for identifying problems that require district action. A leading practitioner notes, "This project has brought to the forefront the flaws of the district. This is a big plus." An external supporter adds, "The pressure brought to bear through this pilot has revealed weaknesses." A board member indicates, "PFP is a good medium, a way to get out of the water and get on shore." Another board member underscores the importance of making improvements, "During the four-year effort, what takes precedence is changing the way to do things."

Pay for Performance has placed on center stage the importance of institutional priorities, leadership and clarity. Two issues are listed below which demonstrate how leadership change and gaps in district alignment have served to weaken the pilot. These are not presented as criticism of the many people in the district who have put in extra effort to advance the pilot. Rather, they reflect problems that have and will continue to affect the pilot as long as the conditions exist. The challenge for district leadership is to build a coherent organization in support of clearly articulated priorities, including Pay for Performance. Recommendations to aid in this process are presented in Chapter X.

Iowa Test of Basic Skills/Control Schools

As previously discussed, the pilot is designed around three approaches to pay for performance. Approach One is based on changes in scores on the Iowa Test of Basic Skills. Approach Two is built around teacher-developed criterion-referenced tests or other teacher-developed measures. This approach has come to be interpreted frequently as criterion-referenced tests which, in many cases, are scored by teachers. Approach Three is built around professional development, focusing on teacher acquisition of skills and knowledge. A requirement to link some form of student achievement to this professional development approach was added during the latter part of the pilot's first year.

The intent of the pilot was to measure changes at these schools as compared to schools not participating in the pilot. Consequently, a clear requirement of the study was to establish control schools so that these kinds of comparisons could be made.

Prior to the pilot, the district required all schools to take the ITBS. It was the most widely used and easily understandable assessment in the district. The district also had longitudinal ITBS data. However, for the 2000-2001 school year, the district decided to drop the requirement that all schools administer the ITBS, as the state's CSAP was growing in importance. The initial pilot schools were finalized in October 1999, based on faculty votes, well into the pilot's first year. The need for control schools was identified at this time, but control schools were not actually chosen until much later. Given the breadth of district activity underway, identifying control schools may not have been seen as a priority.

Control schools were initially identified in Fall 2000, but were not formally notified either as to their status or the implications of that status at that time. Due to the leadership changes, there was a lack of clarity regarding decision-making authority. The approval of the list of control schools was therefore delayed. Meanwhile, when the use of the ITBS became optional, some schools chose to stop using it. Thus, into the pilot's second year, the district lacked a set of control schools whose test results could be compared to those of the pilot schools.

A series of meetings, letters and discussions followed the identification of the urgency of this problem. The resolution was that control schools were formally identified, and notified that they would be required to administer the ITBS to all of their children (second grade and above), just weeks before that test was scheduled for the Spring 2001 administration. Those schools that had decided to abandon the test, in accordance with district policy, were understandably inconvenienced and angered. This anger was directed primarily at the pilot and those associated with it.

The action to designate control schools and use the ITBS remedied problems that were not anticipated when the pilot was approved. However, the delay in doing so resulted in tensions which could otherwise have been avoided. The district has learned from this experience.

On the face of it, the selection and notification of control schools was not a difficult task. On several occasions, administrators and board members were reminded of and acknowledged the need. The problem at that time, as in other instances, was that the pilot was not sufficiently incorporated into the district's managerial and decision-making structures. This problem was exacerbated by the frequent turnover of leadership.

Teacher Identification Numbers

One of the presumptions of pay for performance is the ability to link teachers to the performance of their students. Since the pilot involves the use of many different measures, this capacity must be comprehensive. Further, it is not enough to simply determine what results a group of students achieved in a given year. Rather, in the Denver pilot, the impact of this year's teacher must be determined based on how much each student in his or her class has learned since the previous year.

The district has expressed interest in examining the results of the pilot based on (1) differentiating between pilot and control schools, and

(2) differentiating among schools, teachers and students by such factors as student demographics or the number of years a teacher has taught, adding further complexity. What may have seemed an easy concept initially—judging teachers on the success of their students—is in reality a serious and complex undertaking for the district. It starts with the ability to link each teacher to his or her students.

Like most districts, Denver's data capacity has grown intermittently in response to requests and reporting requirements. Also, many different people and departments collect data at many different junctures. From CTAC's national experience, Denver is distinct from many districts by virtue of having a greater number of skilled and knowledgeable central staff—knowledgeable about technology, assessment and statistics—than most. Even with this proficiency, however, the challenges of data management confronting the district are significant. The needs of Pay for Performance have stretched the capacity of the district.

The problem of tracking student achievement by teacher starts with the designation of teacher identification numbers. In Denver, these numbers have been generated historically at the school site level, most often by designating a teacher by the number of his or her classroom. Teachers who change rooms, change schools, or otherwise move end up changing their identification numbers regularly. New teachers are assigned numbers held by teachers who retired the previous year. Even within a given year, the system is unreliable for linking particular students with a particular teacher. Over multiple years, the overall system is severely deficient.

As different computer data systems have developed differently in different departments, moreover, even standardizing how social security numbers are recorded (with or without dashes) can cost many hours of labor to resolve. The widely used student record system that Denver now employs provides only a four-digit space for recording teacher identification numbers. This is not sufficient to provide unique numbers to Denver's teachers. None of these problems are unsolvable. Yet each solution requires time and person power, both of which are in short supply. Further, the solutions require that the district address these problems as a priority across departments. They cannot just be an additional assignment placed on top of the many assignments the relevant staff already have. The addition of significant new challenges requires either adding staffing or reordering departmental priorities.

The problem of unique teacher identification numbers was identified at the outset of the research study. Over the course of the first half of the pilot, countless meetings were held and memos issued. The short-term resolution has involved telephoning the pilot schools at three points during the year to double check teacher assignments. This is extremely labor intensive and highly prone to error. It is used only for the pilot schools, and could not possibly work for the entire district.

The departments that have engaged in collecting these data have worked hard at the process, but the institutional imperative to develop a systemic solution has not emerged during the pilot's first two years. In this context, two issues remain clear. First, a pilot which cannot collect the appropriate data will not yield sufficient information. Second, any form of pay for performance—or any performance appraisal system tied to student achievement—must ultimately have a strong data linkage system, including unique teacher identification numbers, as a fundamental component. These problems will remain until there is a clear institutional mandate to address them. Recommendations regarding the development of such a data system are included in Chapter X.

G. Administrator Pay for Performance

In the 1999-2000 school year, the district also created a version of Pay for Performance for administrators. The administrative version created controversy within the district—a majority of the central administrators reached their objectives, while a significantly smaller percentage of principals met theirs. For a time, this resulted in misunderstanding over the core concepts and considerable anger within the ranks of principals and assistant principals.

While this problem had no direct relationship to the pilot for teachers, it had a considerable negative and indirect effect. For example, controversy arose as the Design Team was trying to encourage secondary schools to join the second year of the pilot. This made a difficult recruitment task even more difficult. The program for administrators has since been terminated. However, it left a residue of negative impressions of Pay for Performance. Also, to the extent that some schools may be disinclined to join the pilot, it has added burdens to the already significant workload of the Design Team.

H. External Influences—CSAP and Underperforming Schools

Like any other initiative, the Pay for Performance Pilot is moving forward in the context of state and national activities. These activities, and their attitudinal underpinnings, have affected perceptions and understandings of the pilot across the district. The following represent a few of the major external influences that have influenced attitudes regarding the pilot.

CSAP and the Report Card System

CSAP is the major statewide assessment of student achievement. It is part of the growing national trend in which the states are attempting to promote educational accountability. The results of CSAP are based on individual and group performance at a particular point rather than on individual student growth. Accordingly, urban communities have tended to perform poorly on the examination. As Colorado's largest and capital city, Denver receives significant media attention. With headlines proclaiming the number of schools rated "D" or "F" during the first year of the report cards, or "low" or "unsatisfactory" during this most recent year, and with the state indicating various penalties for low performance, the CSAP approach to accountability has dominated public discourse.

The Pay for Performance approach to accountability is focused on a teacher's contribution to the learning of individual students in his or her classroom in a given year. Yet teachers may help their students to advance considerably in an urban school, meet their objectives, and still the school may receive a low rating from the state. Thus, while the Denver and state initiatives share a common interest in promoting student achievement, they are constructed very differently.

The differences between pay for performance and the overall CSAP approach have confused discussions about teacher accountability and ways to look at student learning. Said one teacher, "The focus on results has not changed as a result of PFP, CSAP is driving that." Another said, "The District should look at student growth as gains, not according to a score on CSAP." Some teachers indicated that they voted to participate in the pilot because they feared that anything the state would produce would be much worse. Other teachers believe that anyone who teaches at a lower performing school will end up being punished because of that. This would be a disincentive for the best teachers to want to teach in those schools. Parents, particularly those whose children attend the lower performing schools, often confuse Pay for Performance (if they are aware of the initiative) with the state's rating system, which many consider punitive and counter-productive. One principal expressed it this way, "Our job has to be to look for a correlation between CSAP and tests used at schools to measure progress. CSAP is the standard we are being held to, and while we must meet our PFP school objectives, we can fall behind in CSAP scores."

Legislation and the Related Climate

A number of bills regarding school accountability and teacher compensation have been proposed during the past two years. Even when pending, teachers and administrators indicate that these legislative initiatives shape the climate for the pilot.

Such legislative initiatives are occurring at both state and national levels. For example, Colorado Revised Statute, CRS Section 22-7-607.5 (formerly known as SB 00-186, Section 19), sets forth a grading system for schools based on their CSAP performance. The recently passed SB 01-098 establishes a teacher incentive program. This legislative focus on issues of teacher and school accountability, in concert with the national educational proposal requiring annual testing and the flurry of other proposals, amendments and actual laws, has created a context of considerable confusion around a pilot that is, in itself, complex. This context has diluted the focus on the pilot. As one board member said, "The climate has changed since we started. With schools being graded…[DPS] will be under the microscope."

I. Organizational Support and Alignment

Design Team

In the contract, the composition of the Design Team was based on representation, rather than functionality. This is common for initiatives growing out of negotiations but is not ideal for purposes of project management. Simply put, the Design Team was not initially structured for management functions. This required the best efforts of Design Team members, coupled with pilot modifications, to improve this situation.

The Design Team faced a series of serious challenges from the very start of the pilot. When school opened in Fall 1999, the Association representatives had been appointed but management had made two temporary appointments of senior administrators. As the year progressed, the district appointed an elementary school principal and a middle school assistant principal. This slowed the start-up of the pilot. With all four appointees on board, the challenges were immediate: team leadership was not designated; there were no operational guidelines or decision-making structure; and there was a critical need to bring schools into the pilot. There was also little time to address these issues.

The Design Team's first year was difficult. Without a designated leader or a vehicle for resolving internal conflicts, even small disagreements could slow or stymie progress. In addition, the Design Team lacked the clear authority needed to establish quality standards, move central departments or require actions from the participating schools. Further, because the pilot existed outside of the district's management structure, Design Team members initially lacked formal mechanisms for obtaining information, communicating needs or interacting with other departments.

Establishing the Design Team generated both intended and unintended consequences. The Design Team was a clear reflection of the collaboration between the Board of Education and the Association. It also showed that the pilot was going to have a special visibility within the district. These were intended and positive consequences. However, an unintended message was delivered to, or perhaps interpreted by, central administrative departments and schools that the Design Team, not the departments, was responsible for ensuring the implementation of the pilot.

The Design Team's role is both catalytic and direct. While the Design Team provides training and works closely in the school on pilot implementation, it can only move the pilot forward with the help—and ownership—of relevant district departments. For example, the collaboration between the Design Team and the departments of Assessment and Testing, and Technology Services produced the Online Assessment Scores Information System (OASIS), an intranet vehicle for delivering student achievement data to the classrooms. However, if the pilot is perceived as being separate from the departments' primary areas of responsibility, it becomes significantly more difficult to make progress. During the initial two years of the pilot, problems related to leadership and organizational imperative have frequently made the tasks of implementation more difficult. They will continue to do so until priorities and direction for the pilot and for the district as a whole are reconciled.

Due to these constraints, advancing the pilot has required an unusually high level of resourcefulness on the part of several parties. First, even when it has generated some tension with the administration, the Board of Education has persisted in emphasizing the pilot as a priority. Second, the Design Team has been rigorous in identifying needs and developing relationships with key central units. Indeed, the pilot would not exist today without the vigilance of the Design Team. Third, even in the absence of cogent leadership mandates, numerous departments have operated on an increasingly cooperative basis with the Design Team. Most notably, the department of Assessment and Testing has played and continues to play a major role in the pilot. Staff members from Curriculum and Instruction, Planning, Technology Services and many other units have also worked with the Design Team to accomplish needed tasks. These efforts have advanced the pilot even in the midst of leadership changes, uncertain priorities and the pile-on effect of adding to workloads without reconfiguring departmental priorities.

J. Modifications

As problems have emerged which have affected the pilot, the Board of Education has stressed the importance of Pay for Performance as a core district priority. The challenge has been to ensure that the board's policy priority becomes an operational priority for other levels of the district. Towards the end of the pilot's first full year, and continuing into the second year, several structural changes were made to strengthen the pilot. These changes, undergirded by the tenacity of the Design Team, have helped to advance the pilot.

Design Team Leadership

A key change was to designate a Design Team Leader, empowered to organize priorities, make and monitor assignments, and resolve any team disputes. This change has helped the Design Team to function with a stronger, unified voice. In essence, team members have increasingly functioned as ambassadors of reform.

Administrative Mechanisms

The district also created mechanisms within the administration to strengthen the pilot. Their impact has varied.

The district needed a vehicle to increase the priority status of the pilot and the related district supports. Accordingly, the role of pilot champion was established. The champion was intended to have the ability and authority to cut through issues of turf and jurisdiction, to enable direct responses to the needs of the pilot and to ensure that decisions would be implemented promptly. The intent of the recommendation of a pilot champion was to have one individual who would be the bottom line support for the Design Team and the pilot. However, the district soon designated two champions, one from the operational side of the district and one from the instructional side. This construct was used, with mixed results, during the second year of the pilot. The new superintendent has recently designated one senior administrator to fill this role.

The district also created inter-departmental vehicles in an effort to support the pilot. This included creating a steering committee to lead departmental responses during the pilot's second year. While this structure facilitated communication, it did not translate into a committee that would influence the main agenda of the different departments. As a result, the pilot continued to be affected by a lack of organizational alignment. In addition, central administrators who had legitimate concerns regarding the pilot continued to lack an effective vehicle for addressing these concerns.

While individual departments often responded to requests, many of the steps needed to support the pilot were and are significant undertakings. They require carefully planned inter-departmental action and restructured departmental priorities. This can only be accomplished with forceful leadership. For example, defining the requirements of all the special subject teachers and specialists, and creating a rubric within which they should set objectives, requires collaboration from multiple departments. Efforts in this area have been initiated by the Design Team. During the balance of the pilot, the district will need to continue to build the capacity to respond to such needs. In this context, the new superintendent has recently established a leadership team to ensure that there is a more effective institutional response to these needs.

Issue Identification

Current district leadership is interested in making the district more of a learning organization. Accordingly, based on available data and the related analyses, the following representative listing of issues was presented to Design Team members and district leaders prior to the start of the third pilot year (2001-2002):

- The need for greater integration of pilot practices into district practices and decision-making priorities;

- The tension that exists between focusing teacher objectives on measurement for the purpose of compensation and on specific learning content, and the need to enhance the learning content identified in teacher objectives;
- Issues of assessment that must be addressed during the latter half of the pilot;
- The need to provide more intensive professional development at the school sites;
- The problems of fairness and inclusion regarding special subject teachers, special education teachers, and specialists;
- The need for the pilot to demonstrate an impact at pilot schools if Pay for Performance is to win the support of teachers and the public.

The Design Team and district leaders have, in many cases, begun initiating actions to address these concerns.

K. Summary

The pilot has made numerous advances, but problems of structure and organizational alignment remain. The success of the pilot to date is a considerable credit to Design Team members, past and present, to the schools and teachers who have been able to work within a situation of distinct ambiguity, and to the individuals within the central administration willing to cooperate outside of their normal tasks to support the pilot. Such levels of cooperation, and recognition of the major challenges facing the pilot, have increased during the first two years of the pilot.

One of the most critical strengths of the pilot has been the willingness of the Board of Education and the Association to make mid-course corrections as necessary. The need for greater ownership and alignment is fully recognized by the sponsoring parties. The new superintendent also recognizes the importance of these issues and has initiated steps to move the administration forward. In the pilot's last two years, a pivotal district challenge will be to fully integrate the pilot within district priorities and management imperatives. Recommendations regarding these issues are included in Chapter X.

CHAPTER IX

Summary of Major Findings and Critical Issues

A. Introduction

This report examines the initial two years of the Pay for Performance Pilot. The first year served as the baseline year of the pilot. In drawing together what has been learned by the end of the second year a picture has begun to emerge—partial, in need of more years of data for analysis, but compelling to those interested in the prospects for PFP in Denver. However, it must be remembered that PFP is a pilot, and that it is *four-year* pilot for good reason. The results of the pilot to date present a clearer pathway than previously existed, but there are issues to resolve and steps to take before Pay for Performance will be ready for a vote by the Association and the Board of Education, or for implementation on a greater scale.

This chapter contains a summary of the findings emerging from the data presented in previous chapters. It also includes a discussion of critical issues and questions derived both from the specific findings and CTAC's broader analysis. Many of these issues have already been recognized by the Design Team and district leaders, but they will require further study and resolution.

B. Summary of Findings

Findings on Student Achievement

In looking for effects of PFP on student achievement in the district, the study analyzed two years of student achievement, based on two years of student data: the 1999-2000 school year (the baseline year for the pilot), and the 2000-2001 year. Two years is useful for comparative purposes, but too short a time to indicate summative trends. Data presented in the final report, which will summarize four years of student achievement, will present a more accurate indication of student, teacher and school trends. With this caveat, and bearing in mind that these initial results may change over time, the findings below tend to support the value of pilot activity.

ITBS Results

- Overall, the elementary schools in both pilot and control school groups were generally performing below grade level (average NCE lower than 50) at the start of the pilot. Exceptions were the students in Approach One schools in math and students in Approach Three schools in reading. Excluding the possibly ineligible students, students in Approach Three schools were also performing at grade level in math.
- Between the baseline year and the second year, students in the control group experienced statistically significant increases of just under one NCE in reading and language and a decrease of just over one NCE in math. Students in Approach One schools achieved as expected—the average changes in NCE scores on all three tests were not statistically different from zero. Students in Approach Two and Three schools also achieved the expected increase in reading and language, but saw declines in math achievement of four NCEs and eight NCEs respectively.
- The pilot elementary schools did not perform differently from the control schools in reading and language, and Approach Three schools performed lower than the control schools in math. While it is perhaps not unexpected that the ITBS results for teachers whose objectives were based on criterion-based tests and professional development would show no effect, teachers in Approach One who created objectives to be measured by ITBS also showed no effect.
- Although the pilot middle school started out with lower baseline scores, its students improved more than students in ten of the seventeen control middle schools (five of which had statistically significant losses) during 2000-2001 academic year in reading and language. In language, the gain of 5.5 NCEs was better than fifteen of the control schools by a statistically significant amount. This is impressive, as six of the controls also had statistically significant, but smaller gains.
- The middle school pilot shows improvement in reading and language scores and greater performance increases than the control schools. Analysis of middle school data was hampered by small sample size (one pilot school) and lack of teacher assignments for the control students, but it is reasonable to conclude that the pilot had a positive effect overall, and probably had a positive effect on Hispanic students, aided students, and low achievers as well.
- This past year, the pilot middle school conducted a school-wide staff development program (focusing on writing and vocabulary) which was highlighted frequently by teachers in interviews.
- Spring 2000 mean reading NCE increased as the quality of the objectives increased (as measured by rubric score) from 41 for the lowest category (rubric score 1) to 52 for the middle categories (rubric scores 2 and 3) to 71 for the excellent category (rubric score 4). Also, the percent of children performing at or above grade level at the end of the previous school year increased in a similar manner. The percent at grade level ranged from 33% for the lowest

quality to mid-50's for the middle two groups, and up to 84% for the excellent category.

- Spring 2001 scores were compared to Spring 2000 for the effect of objective quality. The NCE scores are norm referenced and scaled such that a change of zero means that children achieved a year's expected growth in reading relative to the national sample given their baseline reading levels. The mean change in student reading scores was close to zero and sometimes even negative for the Needs Improvement, Partial, and Acceptable categories (rubric scores 1, 2, 3). However, the students of teachers with excellent quality objectives had increases of 2.6 (Objective One) and 1.4 (Objective Two) NCE points on average. This indicates that students of teachers who wrote excellent objectives achieved more than a year's growth, on average, while students whose teachers had less than excellent objectives, averaged the expected year's development and no more.
- With Excellent objectives (rubric level 4), the high quality of the objectives correlated positively with increases in student achievement—whether the objectives were met or not met. This is a particularly noteworthy finding.

CSAP Results

- Pilot school students performed significantly higher in third grade total reading with a mean of 534.98 for pilot school students compared to 527.06 for control school students.
- On average, lower reading scale scores were found for Hispanic students in the pilot schools when compared to other students in their classrooms and for lower socioeconomic status students when compared to other students in their classrooms.
- Pilot school students performed significantly higher in fourth grade reading for Standard 1 with a mean of 567.33 for pilots compared to a mean of 557.50 for control group classrooms.
- Pilot schools had lower mean scale scores than comparisons on the Spanish language version of the CSAP on almost every outcome, and in two cases, Standards 4 and 6, the control school students scored significantly higher.
- The number of years a teacher has been in the district is positively associated with CSAP reading scale scores, with a one-year increase in teacher experience associated with a 0.88 difference in reading averages. It is not clear whether this measures a positive effect on students of more experienced teachers, or a tendency for the experienced teachers to teach in higher performing schools.
- On the third grade version of CSAP, the quality of objectives is not a significant predictor of reading scale scores.
- On the fourth grade CSAP, the quality of teacher objectives and pilot approach were found to be significant predictors of fourth grade CSAP reading scale scores. Of the 36% of variation not explained by student factors, 60% can be explained by the quality of the objectives and 61% by which approach teachers used.
- The quality of teacher objectives correlates significantly with student success, no matter which approach is used. On average, a one-point increase in the quality of the teacher's objectives is associated with a thirteen-point increase in average student reading scale scores.

Findings on Assessments of Student Achievement Used in the District

One of the outcomes of using student achievement as a basis for compensating teachers is that student assessments bear a considerable burden in terms of validity, reliability and utility. Assessment problems and questions that have emerged—not only from the statistician's perspective but also from that of the classroom teacher—are far from resolved and will require considerable attention over the remaining years of the pilot.

- For purposes of measuring a teacher's direct contribution to student achievement in a given classroom over the course of a given year, assessments measuring student growth (providing a pre-post measure) are the most appropriate. Of the standardized tests most widely in use, only ITBS provides this ability. CSAP is moving in this direction.

- Although the ITBS provides the most easily used assessment of student growth, it is the most removed from specific classroom content, and thus difficult to use in setting specific learning objectives. This deficiency can be remedied, to some extent, through the use of ITBS item analysis and clearer alignment of the objectives to the curriculum.
- The two most widely used assessments (ITBS and CSAP) are administered in a relatively uniform pattern, but many discrepancies remain within and among schools, limiting the ability to create cross-school comparisons.
- Usable data for these analyses leave many questions because of implementation discrepancies and other data omissions. In the sample under study, many students had missing assessments or teacher identification numbers.
- Some of the assessments in use in the district are more useful for purposes of classroom instruction than for comparisons of student growth. First, assessment results for young children typically vary widely from one test to the next. Second, many instruction-based assessments, such as the Colorado Basic Literacy Assessments (DRA, QRI) and the Six-Trait Writing assessment, are highly subjective, based on teacher evaluation of student progress. While very useful in the classroom to benchmark student learning, these assessments present difficulties when used for the purpose of compensation.
- Too little data exist for teachers to evaluate the strengths of their students in specific areas at the beginning of the year, the starting place from which to determine needs and measure growth. This is the case for several reasons: the difficulties of providing student achievement data on a classroom basis at the beginning of the year; the lack of standards aligned with appropriate assessments and communicated to teachers; and the need for additional classroom diagnostic assessments.
- Notwithstanding the above, the provision of student achievement data to the pilot schools, and the related teacher understanding and use of data, has increased substantially as a direct result of the pilot. This has occurred through the actions of the Design Team and the initiatives of the Assessment and Testing department and others, including the creation of the OASIS data system.

Findings on Objectives and Objective Setting

Teacher-set objectives have been central to the pilot since its inception. Having identified a tension between objectives set for the purpose of measurement and objectives set as the basis for student achievement, CTAC created a rubric based on student learning as a means to analyze objectives and compare them between years of the pilot, with school plans, with control school goals, and with student outcomes.

A full analysis of objectives is presented in Chapter IV, and links between objectives and achievement are discussed in Chapters V through VII. Additional conclusions related to objectives, including the alignment of curriculum and instruction and the appropriate use of assessments, are discussed below.

Classroom Teachers

- Objective setting improved from the first to the second year of the pilot.
- Teachers who scored highest on the objectives rubric showed higher student achievement results on both CSAP and ITBS.
- Objectives tend to focus on measurement rather than specific learning content, for the purpose of compensation.
- Pilot school objectives are more specific than control school goals, especially with baseline data, but both tend to emphasize measurement more than learning content.

- Teachers should be encouraged to focus on learning content in their objectives, describing assessment and improvements in achievement as an outcome of that content.
- Teachers appreciate the Design Team training in objective setting, but more professional development is needed and wanted.
- Training provided to date has advanced objectives to their current level, but additional training will be needed, especially as content-focused objective setting appears not to have been a dominant district practice.

Special Subject Teachers, Special Education Teachers, and Specialists

- Objectives for special subject teachers, special education teachers and specialists should also be content-focused, but the assessments should be more appropriate to the responsibilities of these positions.
- The Design Team has initiated discussions around the expectations for these staff positions.

Findings Based on Viewpoints and Perceptions of PFP Participants

- Teachers say that they are not doing things differently than in the past. However, they also indicate that some things at their schools have changed—particularly an increased focus on student achievement.
- Teachers also indicate positive developments and improvements in the responses they have received from the Design Team.
- Teachers tend not to believe that the current PFP financial incentives will make them teach any better or try any harder. However, many do believe good performance should be rewarded.
- Teachers are not clear on the purposes of the pilot. Although they strongly support the notion that increasing student achievement is a primary purpose, they are unclear on the ways in which this can or should be linked to compensation.
- By and large, survey respondents believe that fair objectives can be set, but many fairness-related issues remain.
- Non-academic teaching personnel, in particular, indicate a range of issues that concern both fairness and how they fit into an overall plan based on traditional measures of student achievement.
- Teachers tend to believe that conditions at their schools have either stayed the same or improved from the first year of the pilot. Only a few see a decline.
- Fears initially expressed about increased competitiveness have not been realized, for the most part. Most teachers see no increase in competitiveness, and a significant percentage see gains in communication and focus.

Findings Related to Professional Development

In survey and interview responses, teachers indicate a strong desire for more professional development, in such areas as setting objectives, utilizing student data, and developing classroom strategies to enhance student learning. Many schools offer professional development to teachers in a variety of areas, but as these are often locally initiated, the quality and scope of professional development vary considerably from school to school. The district also offers training and professional development to schools, much of which is also available at the discretion of the school. The full scope of professional development activity, the relationship between this activity and student learning needs at the schools, and the adequacy and effectiveness of the training, are not clear.

The Pilot and District Operations

Since a core purpose of the pilot is to improve student achievement, the resultant setting or confirming of standards, curriculum and assessment alignment, professional development, priority setting, and data analysis that are needed for the pilot to succeed are the same tasks that the district must undertake to reach its stated goals. The initial construct of the pilot was developed as discrete from district operations, and the mechanisms set up to implement the pilot, such as the Design Team, were originally outside of the district's administrative structure.

Alignment of instruction: Evaluation of curricular and instructional practices within the district is beyond the scope of this study, but it appears that these practices differ significantly from school to school and confound the teaching and learning issues of the pilot. While state standards exist, and while DPS documents defining and explaining those standards have been published, use of these standards in defining curriculum and instruction appears to be uneven from one school or classroom to the next. Performance on state tests based on state standards is weakened if a common set of standards is not used in aligning curriculum, instruction and assessment. Efforts to link objective setting both to classroom assessment and broader assessment, as well as efforts to assure fairness across objectives, are hampered by this lack of consistency and focus.

Integration of pilot practices into district/supervisor expectations: A separation of pilot responsibilities from department priorities has sent a mixed message concerning the priority placed on a successful pilot by the district.

Communication: Even with considerable effort expended, many pilot teachers are often unclear about the purposes and other aspects of the pilot. Many control teachers are even less clear. A coordinated effort at communicating pilot purposes, the eventual decision that must be made, the findings from the first two years of the pilot, and, as they are identified, potential options for implementation, should be developed. The Design Team and district have launched a comprehensive communications effort to accomplish these purposes.

C. Discussion of Critical Issues and Questions Arising from Findings

District Capacity and Priorities

District Ownership and Management Involvement

A clear finding arising from interviewing central administrators is that they understood the Design Team to be in charge of PFP and did not, at the launching of the pilot, assume responsibility for the success of the pilot. Though many central administrators became increasingly involved in and supportive of the pilot during the first two years, that involvement still fell short of assuming responsibility in many instances. As one example, supervisors of pilot schools generally did not discuss requirements of the pilot with principals during the initial two years. They did not hold principals or schools responsible for conducting pilot business, nor convey to schools the importance assigned to the pilot by the Board of Education. This is an area which the current administration is seeking to address.

Professional Development

Since the time of the pilot's inception, the district has lacked cogent leadership in the area of professional development. Accordingly, while the district engages in a substantial amount of training and professional development, it is not clear whether this is appropriately focused, of consistently high quality, or effective. Professional development is needed to enhance the pilot, and far greater professional development capacity will be needed as the district proceeds with Pay for Performance.

Data Capacity

The study has noted previously that Denver's technical proficiencies exceed those of many districts where we have worked. Nonetheless, it is clear that Denver's systems were not designed with teacher accountability or compensation that is directly related to student performance in mind. The core challenge is to build a linked data system from existing and new components within the district to support a pay for performance. DPS has the internal capacity to design such a system largely from within, but relevant staff members must be given the appropriate charge and release time to move in this direction.

Curricular and Instructional Alignment

Evidence from the sites suggests that there has not historically been a strong central core to the district's academic program—standards, curriculum, instruction, and assessment. Rather, it appears that guidelines have been issued, via well-developed curricular outlines produced by the administration, but that implementation of much of the curriculum has been left substantially to local discretion. An example of this approach is reading. A myriad of different reading programs are in use at different schools within the district. Some may be strongly linked to common standards, but the evidence suggests that not all of them are. It has not been a part of the study's charge to conduct an audit of curriculum and instruction in Denver. Nonetheless, the evidence strongly suggests a level of site autonomy around instructional *content* that may undercut Pay for Performance, and that may also undercut the district's goals for student achievement.

Public Support

Two perceptual issues are particularly important with regard to public support. First, the public must perceive that any funds being expended in support of the pilot are serving a positive purpose. If PFP continues to show a positive correlation with student outcomes on the ITBS and CSAP, this will not be difficult. Overall, the district will have to be able to show how additional compensation is merited in terms of student achievement and how the schools benefit from pay for performance.

Second, parents seem to have limited knowledge about pay for performance, but they are concerned with the kinds of objectives that teachers may be paid to achieve. The involvement of parents in some form in the pilot, perhaps through the school planning process, would likely create another constituency in support of pay for performance.

Issues of Assessment of Student Progress

Assessment of student progress is the point of connection between student performance and teacher performance—the linkage around which the pilot is constructed.

Because there are many questions and disagreements concerning academic assessment, and because the pilot seeks to find a way to link teacher compensation to student achievement through this vehicle, the district is obligated to approach the use of assessments both carefully and thoroughly, considering the strengths and weaknesses of both the assessments themselves and the analytical approaches used to understand their results. Below are some of these assessment issues.

Challenges for an Assessment System

The challenges for individual assessments, and for a system of assessment, include the following:

Student Growth: An individual teacher cannot be held accountable for the starting places of his or her students at the beginning of the school year. To the extent that a teacher is judged on the basis of student achievement, he or she must be judged on how much the student has progressed from that starting place. Thus, each assessment must measure student growth, not just absolute student achievement. If an assessment does not meet this challenge, it penalizes teachers for students who start further behind. CSAP was not designed to accomplish this purpose. It has been indicated that the Colorado Department of Education is working to scale CSAP assessments such that they can be used to measure the progress or growth of individual students.

Alignment of Curriculum, Instruction and Assessment: Assessments must reflect the goals of the course or level within the subject that is being taught. Teachers have commented that student assessments are measures of student performance, not teacher performance. This is true to the extent that assessments do not reflect what is being taught. The need for alignment of standards, curriculum and instruction is discussed elsewhere in this report. The particular point about assessments within this alignment is that they only measure the contribution of a teacher toward a student's achievement if they reflect what the teacher has taught. The most useful assessments for PFP, therefore, are those that directly relate to the instruction provided by the teacher.

Different Assessments for Different Purposes: A distinction must be made between assessments which gauge a teacher's contribution to student achievement and assessments used for the purpose of comparability. An assessment that closely matches what a teacher has taught is best for measuring that teacher's contribution to learning. If teachers are teaching different material, it will not be useful for comparative purposes. A more general norm-referenced test like ITBS is designed for comparisons.

Reliability and Validity: Assessments must have proven reliability and validity, both statistically and perceptually. This involves issues in the development of the tests, their link to the curriculum and instruction (what is taught), and the ways they are implemented.

Unintended Consequences: Assessments should not narrow teaching to a set of facts, nor lead to negative behaviors by, or consequences for, teachers or students. Curriculum and instruction should drive assessment, not the other way around.

Frequency and Consistency of Implementation: Assessments that are to be used for comparative purposes must be conducted at all schools under similar conditions—given at the same times, to similar students, following the same procedures. Assessments that are not implemented in this way are less valid for comparative purposes.

District Data Capacity: Since teachers are expected to construct objectives based on identified student needs, the data system must have the ability to provide information to teachers on their students. Achievement will increase the most (students and teachers will be the most successful) if teachers address the areas of each child's greatest need. Assessments that are linked to instruction and are capable of providing analyses of specific areas of student need will substantially increase the likelihood of increases in achievement. Significant progress has already been made in this area, but the district does not have an overall data system capable of supporting a large scale pay for performance effort.

Statistical Issues

Statistical Validity at the Classroom Level: Though assessments that produce quantitative results are favored across the country, the conclusions being drawn from some of these assessments raise questions of statistical validity. While valid across larger school systems or states, the differences between individual children, or even individual schools, take on increased importance as the numbers of students get smaller. This statistical fact has led to a multitude of problems with assessments nationally, and has plagued the pilot as members of the Design Team and others have looked for fair and valid approaches to assessment. If tests such as the ITBS and CSAP are to be used to assess student achievement at the classroom level, new ways of analyzing and disaggregating the data must be explored.

Alternative Statistical Methodologies: To meet this challenge, CTAC has introduced several different statistical methodologies to the problem of measuring achievement at the classroom level. Internal and external experts have been asked to apply the techniques of the value-added approach to assessing achievement, as well as different approaches to statistical modeling: individual growth modeling and hierarchical linear modeling. The results to date of several of these approaches are presented in Chapters V and VI. These processes are ongoing, and will require additional years of data to produce definitive results.

Alternative Approaches: It is also possible that none of the statistical methodologies will be sufficiently accurate and clear that the parties will be comfortable in their high stakes use. One possible approach to addressing these statistical problems might be considered. This would involve using several years of teacher data as a basis for making a multi-year decision regarding performance. For example, if three years of students taught by one teacher were analyzed together as a single three-year example of teacher effectiveness, a number of the questions of validity arising from the small numbers of students and the impact of different classes would be resolved. Another possibility, advanced by several members of the DPS administration during the first two years of the pilot, is to use tests such as ITBS to establish group goals and bonuses (by grade or by whole school). The Joint Task Force on Teacher Salary should consider this topic as part of its agenda.

Appropriate Test Implementation— The Six-Trait Problem

Most tests that produce numerical outcomes can be analyzed statistically. This does not necessarily mean that the results have validity, however; valid test results depend on what the tests measure and what the numbers mean. The case of the 6+1 Trait Writing program, commonly referred to as Six-Trait, illustrates this problem, but many of the assessments used in the lower grades present similar problems with regard to Pay for Performance.

According to its developer, the Northwest Regional Educational Laboratory, the 6+1 Trait Writing program was developed as "an instrument that would provide accurate and reliable feedback to students and teachers that would guide instruction" around writing. Its primary purpose is to guide classroom instruction of writing.

In Denver, the Six-Trait Writing test is the second most widely used test of achievement in the district. It is administered from grades three through 10 to students at all schools twice a year. The results are collected and records kept, but analysis is not typically undertaken beyond the classroom level.

The value of Six-Trait as an instructional tool lies in the emphasis of the measurement on attributes of writing that are deemed most important, on the use of actual student essays rather than question-based substitutes (multiple choice), and on the direct usefulness of the test/process by teachers in classroom instruction. Students and teachers review essays according to a rubric, and assign particular scores. The students can rewrite essays, review each other's essays, or work with the teacher. As a teaching tool, especially when used in conjunction with a good professional development program, the Six-Trait approach seems valuable.

With regard to Pay for Performance or cross-district comparisons, however, Six-Trait presents three sets of weaknesses. The first is that the best and most common usage of Six-Trait assessment for instructional purposes is for the teacher to score the essays of his or her own students. This enhances instruction. It also introduces both bias, in that the teacher knows the students, and a question about the reliability of the results, in that the teacher is awarded a bonus based on his or her own scoring system. When higher stakes are introduced, the need for objectivity, comparability and accuracy will increase.

The second weakness is that the scoring of any "open-ended" test is labor and time intensive, involves training, and significantly increases the possibility of human error. The difficulties of teacher bias would be reduced by having teachers exchange essays between schools, for example, as reportedly happens between some of the schools in the pilot. Even if this were done, however, evaluating student essays according to a rubric is still an inexact science. The problem of different teachers awarding different scores to the same essays—the problem of inter-rater reliability—remains. Training is required, and is provided by the district, but training reduces rather than eliminates the problem of inter-rater reliability.

A way to assure inter-rater reliability is to have multiple readers. A common approach is to have two trained readers read and rate each essay and to use a third reader when a discrepancy between the first two readers occurs. This is an effective way to assure comparable results, and would make Six-Trait and similar assessments more valuable in a PFP program. But this approach also demonstrates the third weakness associated with the use of such assessments for comparative purposes: any effort to assure that a given score on Six-Trait carries the same meaning across all classrooms and schools will substantially increase the difficulty of implementation. Having three readers for student essays across the district, once or twice per year, is a substantial commitment of time on assessment.

Some districts have chosen to spend that kind of time on particular tests, and have found the results valuable, but this is a decision which must be made carefully.

Currently, the administration of Six-Trait in Denver is left largely to individual schools. In some schools, student essays are collected and scored by a group of trained teachers. In others, arrangements are made to exchange essays from one school with essays from another, so that teachers don't know the students. Essays are not always scored blindly, which is to say that in some instances teachers know the children who wrote the essays and in some instances they do not.

These practices constitute problems when higher stakes are tied to individual assessments. These same problems also apply to other developmental assessments that are generally teacher scored within each classroom. The challenge of integrating assessments, learning objectives, and teacher compensation will be pivotal for the Design Team, relevant departments and the Joint Task Force on Teacher Salary in the final two years of the pilot.

Appropriate Test Usage—The Student Assessment Problem

Volumes have been written about the uses to which various assessments of student achievement are and should be put. The issues raised cannot be fully identified or discussed in their complexity in this report. However, the pilot cannot be successful unless the limitations and caveats that attach to assessments are at least generally understood. Several of these issues and limitations have been identified:

- The use of assessments at the classroom level creates issues of statistical validity that have not yet been fully addressed.
- The assessments most appropriate for setting objectives and supporting instruction are less useful in comparing student achievement across students.
- In measuring a teacher's contribution to learning, assessments must closely parallel the curriculum and instructional content—what the teacher has taught.
- Fair use of assessments in any high stakes setting, including for purposes of compensation, requires careful attention to implementation, which may be cumbersome and expensive.

Three additional issues with quantitative student assessments, such as CSAP and ITBS, must also be identified and discussed. First, the more generally useful a test is for comparing student growth, the less useful it tends to be for instruction. The Six-Trait example has already been provided. Six-Trait is a highly useful tool for teaching writing, but a difficult tool to use for broadly determining either student achievement or teacher performance. CSAP is more closely linked with state standards, which should eventually drive an aligned system of curriculum and instruction in Denver. But, like many such tests, it is tied to broad definitions of proficiency by grade level, and reported as a relative level of achievement between schools. The CSAP rankings do not currently factor in demographic differences between schools and communities.

Put another way, the closer an assessment is tied to meeting specific standards at specific points in time, the less it is often tied to measuring student growth. Since student growth is the best approach to assess teacher impact on student achievement, this presents an analysis problem. The ITBS is primarily a comparative test, a test referenced against national norms of performance and capable of charting student growth. For comparative purposes, this is by far the best comparative measure in Denver's arsenal of assessments. The problem, however, is that it is the farthest removed from what teachers actually teach, making it a weaker measure of teacher proficiency. No single test in Denver currently provides a close link to student growth, a close link to what a teacher has taught, and broad comparability.

Second, statistical analyses have shown repeatedly that the greatest amount of variation among students on any given assessment has to do with student factors, and that these variations are indeed large. States that have awarded funds to schools achieving the greatest growth have found that many of those schools often do not qualify in successive years. Nationally, it is estimated that five percent of the top-performing schools in most states, measured in terms of growth, are top-performing three years in a row. For example, Kane and Staiger of the Brookings Institute suggest that, while more than half of states reward or punish good or bad schools, "...most of these [assessment] systems have been set up with very little recognition of the strengths and measures that they're based on." Further, they found "…that between 50% and 80% of the improvement in a school's average test scores from one year to the next was temporary and was caused by fluctuations that had nothing to do with long term changes in learning or productivity."[1]

This finding does not indicate that these schools are good or bad, or even that they are inconsistent in performance. It simply demonstrates the great variability between different children, and demonstrates the limited reliability of a single test score, no matter the quality of the test. To date, state programs have yet to overcome this variability. Thus, the question of how much gain can be expected from a child or a classroom is difficult to answer statistically. Through its modeling and value-added approaches, CTAC is generating models or ranges of expectation. But it is likely that multiple measures and/or multiple assessments (over several years) will be necessary.

Third, there is the question of unintended consequences. Research has shown that:

- Scores on any test tend to go up as teachers and students become more familiar with the test.
- Drop out rates and other undesirable outcomes often increase before particular high stakes versions of tests. For example, where states require passing a tenth grade test to attain a high school diploma, increases in the dropout rate, which tend to skew towards poor and minority children, often appear in ninth grade.
- The focus on particular tests tends to increase as the stakes for that test increase. Up to a point, this may be positive, but repeated instances of teachers who abandon curricular units to focus on test-taking skills, and even of teachers and administrators who falsify records, show that the negative effects of tying high stakes to tests must be considered.
- Focused high stakes assessments sometimes result in a narrowed curriculum. Science, music, art and history have all suffered in some schools where reading and math, or test-taking, have been overemphasized.
- High stakes tests often create high anxiety in the classroom. Policy makers sometimes dismiss the effects of such anxiety, but parents and teachers are far less inclined to do so. Parents in many states and communities have rejected placing young children in stressful situations. Further, teachers will not be attracted to an environment that is perceived as high stress, although they will be interested in a system that is seen to reward high performance. The challenge for the pilot and district is to encourage and reward high performance in an atmosphere perceived as supportive and fair.

With regard to high stakes testing, which includes tests used for significant decisions about students or teachers, the non-partisan National Research Council has reached several relevant conclusions:

> "The important thing about a test is not its validity in general but its validity when used for a specific purpose. Thus, tests that are valuable for 'leading' the curriculum, or holding schools accountable are not appropriate for making high stakes decisions about individual student mastery unless the curriculum, the teaching, and the tests are aligned…."
>
> "Tests are not perfect. Test questions are a sample of possible questions that could be asked in a given area. Moreover, a test score is not an exact measure of a student's knowledge or skills. A student's score can be expected to vary across different versions of a test—within a margin of error determined by the reliability

> of the test—as a function of the particular sample of questions asked and/or transitory factors, such as the student's health on the day of the test. Thus, no single test score can be considered a definitive measure of a student's knowledge...."
>
> "An educational decision that will have a major impact on a test taker should not be made solely or automatically on the basis of a single test score. Other relevant information about student knowledge and skill should also be taken into account."[2]

As a consequence, the Council recommends that, "Tests should be used for high stakes decisions about individual mastery only after implementing changes in teaching and curriculum that ensure that students have been taught the knowledge and skills on which they are being tested."[3]

Given all of these difficulties and caveats regarding testing, how does the pilot proceed? A summary of points to be considered as the pilot moves into the development of Pay for Performance options is provided at the end of this chapter. In Chapter X, this report provides specific recommendations for district and pilot action. During the second half of the pilot, these and other issues of assessment should be explored. Many of these explorations have already begun; others should begin soon. The promise of teacher pay for performance is not just in incentives, but in building a system of education that systematically focuses on, and rewards, taking those steps that increase student achievement and teacher effectiveness.

Issues of Fairness

The issues of assessment described above lead directly to a discussion of fairness. Not only must a system of teacher compensation be fair, it must be perceived as fair by the majority of teachers and by a majority of the public. Fairness is an elusive goal. Certainly, all parties must accept that total fairness can never be possible. However, there are significant areas of fairness that have been identified by teachers and administrators, and that must be addressed in the development of the final Pay for Performance options.

Teacher Support

Teachers require information and training of high quality if they are to make changes in their classrooms. Through the efforts of the Design Team and members of DPS departments of Technology Services and Assessment and Testing, the OASIS system has been developed to bring data to the sites. This system should be expanded, and additional training should be provided to maximize this new level of access to student information. Also, the data provided through OASIS should tell teachers what they need to know about their students, and teachers need to know how to make sense of the data, or the system is still not complete. To the extent that teachers are expected to change, the necessary information and supports must be in place. These will most often include training, information, and time. Provision must be made in the final Pay for Performance options to provide teachers with the support they need to succeed. If teachers are expected to behave differently, how the district behaves in supporting teachers (as well as in holding them accountable) must also be different.

Special Subject and Special Education Teachers, and Specialists

Teachers in these categories make up a substantial portion of Denver's teaching force. A careful reading of objectives makes clear that the approaches the special subject and special education teachers, and specialists such as psychologists and nurses, have taken to writing objectives has varied widely. Some have developed their own approaches, others have been guided by school requirements. Many have struggled, especially in the context of the pilot's three formal approaches, to figure out how they fit in. In some instances, for example, physical education and art teachers have written objectives related to improving reading scores on the Iowa Test of Basic Skills. In other instances, objectives for these teachers have related to their own subject areas.

Survey and interview results indicate that these problems have created some consternation for teachers in these categories. They have struggled to make objectives meaningful, to relate them to school goals and their specific approach, and to relate them to their own professional responsibilities. Some have found a formula for writing these objectives that makes sense to them, but many have not. To the extent that the writing of objectives becomes more high stakes, as it may if more than bonuses are involved, concerns about fairness will rise considerably.

Through the initiative of the Design Team, the district has begun to pursue a course of action regarding teachers in these categories. That is, it has initiated a review of the professional requirements of these positions such that guidelines or a curriculum, like those for regular classrooms, are developed and promulgated. It has been indicated that various working groups have made varying amounts of progress in these areas, although it is not clear that these efforts cover all of these positions.

Teachers and other professionals are hired to contribute in areas of need, and should be judged on their contributions in those areas. While a gym teacher can put signs up that encourage students to read, the reason for physical education is not to enhance reading, and the contribution the teacher makes should be judged on measures appropriate to the position.

Differences Among Classrooms and Schools

Among the concerns addressed by teachers in surveys and interviews is the question of whether differences between classrooms might skew the results of a pay for performance system. These concerns, and the differences they address, take many different forms. Some have expressed concern that teaching at a low performing school, where children often start at lower levels and come to school with health or social issues, penalizes them. Some teachers have indicated that differences between classrooms at individual schools, or differences between children within the same classroom, make pay for performance inherently unfair. For example, if one classroom has one or two children with special needs, does that create a classroom context where the teacher is at a disadvantage in Pay for Performance? If so, or if it is perceived to be so, does that discourage teachers from taking responsibility for more challenging or lower performing children?

Some of these concerns are or should be addressed by the pilot's focus on student *growth* rather than absolute or relative attainment. Thus, teachers at the lowest performing schools should have equal opportunity for additional compensation if they advance their students as far as teachers at other schools. But such a model presents its own problems. The district will need to consider that a growth model may mean that many teachers receive additional compensation under Pay for Performance even though the school may be considered to be low performing by the state. Where this is the case, careful explanation to the community will be required.

At the classroom level, the district should emphasize differentiated instruction as needed, including providing training for teachers in working effectively in classrooms where achievement levels differ widely, as a core component of the professional development program. To the extent teachers have not received such training or desire more, this training will advance the pilot by increasing the likelihood of student success across the district.

Fairness and Subjectivity

Finally, the pilot has been constructed around objectives set by teachers with a principal's approval. A part of the reason for this construct was to assure teacher and principal discretion in identifying the differing needs in different classrooms and addressing them appropriately. Whatever the final form of Pay for Performance in Denver, an element of subjectivity will always be required to ensure the flexibility that fairness requires. A component that should be developed in the final two years of the pilot is to provide training for principals, not only through the

Design Team but through the district's area structure, on classroom observation, appropriate objective setting, and the uses of assessment data. At the same time, teachers must realize that flexibility leads to greater subjectivity, and a balance must be struck. Teachers and the public must be assured that the system is generally fair, providing both guidelines and flexibility, if the pilot is ultimately to succeed.

Specific recommendations for district action are presented in Chapter X. These are based on the findings described throughout this report.

CHAPTER X

Issues and Mid-Point Recommendations

The Denver Public School district is well positioned for significant progress. At a time when many districts feel pressured to undertake quick fixes, the Board of Education and the Association are choosing another, more substantive path. They are undertaking one of the most courageous and far-reaching experiments in American public education—the Pay for Performance Pilot. This initiative is unusually innovative in that it seeks to create a direct linkage between student achievement and teacher compensation. The accomplishments to date are distinct. At the mid-point of the pilot, the district must now build on these accomplishments so that an emphasis on results pervades both the pilot and the district as a whole.

Pay for Performance has brought many of the core elements of systemic reform into focus. The pilot is one of the highest priorities of both the Board of Education and the Association. The partnership of these two parties—on behalf of student achievement—is unique. The new superintendent and leadership team are incorporating the pilot's challenges of implementation into their discussions and exploring ways of changing how services are delivered to school sites and classrooms. The Design Team is serving as a catalyst for taking lessons learned to scale. The participating schools are seeing the value of strengthening classroom objectives. Pay for Performance is showing the potential to be a true pathway for results. However, many serious challenges lie ahead.

Further progress is needed on several fronts. As in many districts, there is a significant need to align instruction, assessment and classroom pedagogy. For PFP to move forward, standards of achievement for all students and teachers need to be identified and communicated to key constituent groups. A plan for implementing these standards must be developed by central departments in close collaboration with the Design Team, school site administrators, teachers and parents. These all need to be supported by an organized capacity which integrates and makes accessible student and teacher data. These efforts must be supported by reordered departmental priorities, reallocated resources, and a competitive compensation system for teachers—the front line service providers. The hub supporting these efforts should be a relevant and timely system of professional development. These challenges are significant, but they can be approached incrementally and systematically during the second half of the pilot.

This report has identified the impact of the pilot's initial years and a range of salient factors which have contributed to that impact. In this context, there are several concerns which shape the climate for implementing Pay for Performance and influence the district's efforts to focus on both results and student achievement. As Denver practitioners examine the potential of Pay for Performance, they are focusing on three areas of concern related to the practicality of linking student achievement to teacher compensation: fairness, measurement and the organizational imperative to move forward.

Fairness is a concern for teachers, administrators, parents and board members. This includes fairness in assessing what is taught, taking into account the differences between classrooms and types of teachers, ensuring that the learning needs and accomplishments of students remain central to the pilot, and ensuring that objectives are set fairly.

In terms of measurement, the concerns focus on the importance of accuracy, precision and appropriateness of the assessments being used. This includes defining student achievement and growth, and assuring that assessments measure that achievement appropriately. Further, practitioners stress the importance of identifying classroom or student expectations for every professional, including special subject teachers, specialists and special education teachers, and identifying assessments relevant to these expectations.

The organizational issues focus on the challenges of implementation. Building on the pilot's progress thus far, both practitioners and community leaders are seeking sustained district priorities—including the focus on addressing the challenges of assessment, student achievement objectives, professional development and the overriding issue of aligning standards, curriculum, instruction and benchmark measures.

Over the next two years, moving forward will require a set of focused actions that address the core challenges identified through the pilot so far. The Board of Education and the Association have already moved the national debate on accountability to the level of a landmark experiment. In this context, one of their greatest strengths has been their willingness to make mid-course corrections to advance the pilot. Midway through the pilot, there is a need for further sustained action.

A summary of recommendations follows. These address issues described throughout the report and highlighted in Chapter IX.

A. Recommendations

Issue One: *Objectives*

OVERVIEW

The objectives are the linchpin of the pilot. They serve as the link between what teachers teach and what students learn. As such, they are the nexus between pay for performance and the actual teacher practices that improve student learning.

The study has found that higher quality teacher objectives (rubric score 4) were associated with higher performance by students on both the ITBS reading and on the CSAP fourth grade reading. These objectives are content-focused. They state clearly what students will learn, demonstrate high expectations for students and include appropriate assessments.

Based on findings, the students of teachers who scored 4 on the rubric achieved, on the average, more than one year's gain—whether those objectives were met or not. This indicates a strong likelihood that fully developed instructional objectives will help move the district toward the goal of improving student performance. Clearly, there is a challenge and opportunity ahead for the Denver Public Schools.

RECOMMENDED ACTION

- *Make learning content the focus of the objectives.* In collaboration with the Design Team, this should become a high priority for the district's area offices, department of Curriculum and Instruction, and department of Assessment and Testing. It requires linking the objectives, instructional materials and assessments with the district's curriculum standards and the students' learning needs. Particular attention should be given to satisfying both of the purposes of PFP objectives: clearly addressing learning content to improve student achievement, and clearly identifying results for purposes of compensation.

- *Provide teachers with more support and options in objective-writing.* The challenge of developing high quality learning objectives is a departure for the district and difficult for some teachers. Teachers will benefit from further support, including the possibility of selecting or adapting from a menu of objectives based on district standards. As they hone their skills, and as the curriculum is further aligned, they can more easily make the transition to developing individual content-focused objectives.

- *Address the fairness issue related to special subject teachers, special education teachers and specialists.* Standard measures of student achievement are not appropriate for all practitioners. Special subject teachers, special education teachers and specialists such as librarians, nurses and counselors need more customized assistance to link their performance to student achievement.

- *Establish a direct relationship between teacher objectives and the school improvement plan.* Both individual teachers and the school community should share the same overall goals. This includes helping schools create plans that are substantive, focused on specific learning goals identified for students at the school, and implemented and evaluated annually.

IMPACT

The objectives are the key to the pilot. Mid-point findings indicate that there is a positive correlation between the quality of teacher objectives and the resultant achievement gains by the students. Building on accomplishments to date in objective setting will significantly strengthen the pilot goal of increasing student achievement.

Issue Two: *District Data Capacity*

OVERVIEW

A teacher performance system that is based, in part, on student achievement must start with the ability to directly link information on teachers, students and classes. This study has noted that Denver's data and analysis proficiencies exceed those of many districts. Nonetheless, it is clear that Denver's systems were not designed for purposes of teacher accountability or compensation.

The core challenge is to build a linked data system from existing and new components within the district to support the pay for performance purposes of student achievement and teacher accountability. The district has the internal capacity to design such a system largely from within, but must be given the charge, management imperative and the priority for staff time to accomplish this.

RECOMMENDED ACTION

- *Build an integrated data system.* This should be based on needs identified through the Design Team and other key departments. This means establishing linkages among data systems for assessment, planning, and human resources, and building each unit's needs and requirements into the system from the ground up. Particular emphasis should be placed on meeting the data requirements for classroom instruction and school planning purposes of teachers and site administrators. There also needs to be a mechanism to identify and resolve emerging inter-departmental data support issues.

- *Assign unique teacher identification numbers to all teachers.* These should follow the teachers throughout their careers in the district. They should also be used at the school level to generate class lists and other teacher-specific documentation. Data systems will need to be changed to accommodate the additional data fields needed to link data sets across departments.

- *Establish and implement uniform expectations for the administration of major district tests.*

Assessments that are used for comparative purposes need to be handled more uniformly than those that are used primarily to aid classroom instruction. Training should be provided to ensure consistency in practice.

- *Broaden district capacity to provide sites with data on student gain.* Using the OASIS intranet system and coordinated professional development as the delivery vehicles, the district should provide teachers and administrators with longitudinal data that is disaggregated at the level of individual student gain. This is a requisite for purposes of objective setting, classroom planning and progress monitoring.

IMPACT

The pilot has underscored the importance of using student achievement data to improve instruction and drive decision-making. Building on the momentum established to date, expanding data capacity will provide the district with far greater ability to support the school sites.

Issue Three: *Assessment*

OVERVIEW

If compensation decisions are to be fair, measurements must be valid and appropriate. The closer a test comes to measuring what is actually taught in the classroom, the more valuable it becomes for measuring teacher performance. A primary consideration for measuring a teacher's contribution to student achievement is how closely and specifically the assessment matches the curriculum and what is being taught. If the test only generally measures what was taught, it can only generally measure a teacher's contribution to student achievement.

The challenges of appropriate measurement are exacerbated by several factors. When assessing teachers' contributions to student achievement, a fair system will need to reduce or eliminate student variation—the fact that teachers have different students in their classrooms, both different from each other in a given year, and different from year to year. Further, there needs to be an emphasis on using tests that measure student growth from one year to the next—that compare the pre- and post scores of the same children. These come closest to accomplishing the goal of measuring individual achievement and make it possible to control for the differences among children that affect average classroom performance.

Of the tests being used in Denver, the ITBS is the only standardized assessment capable of using student growth to measure achievement. Six-Trait uses student growth, but is not standardized and raises issues of inter-rater reliability when administered by the classroom teacher. CSAP comes closer to measuring actual classroom instruction, to the extent that Denver's curriculum is aligned with state standards, but does not yet measure individual student growth.

In summary, the best tests for assessing classroom performance—those that measure specific learning content—are content specific tests. However, such content specific tests are not the best tests for broadly comparing students across many schools or grades, since measuring many students across schools or grades requires greater ability to generalize. General tests (such as ITBS) may work the best for measuring across schools and the district.

RECOMMENDED ACTION

- *Develop a means for using multiple measures at the classroom level.* This will address three key concerns. First, using multiple measures will help address the weakness of any given assessment in actually gauging student achievement or teacher performance. Second, combining the use of different assessments will help solve the statistical problem of small numbers of students in a classroom. Third, using multiple measures will better allow the district to fairly address the need to measure student achievement and teacher performance in a given classroom, and to broadly compare students and classrooms across schools in the district. The subsequent use and implementation of multiple measures should be evaluated by the district.

- *Select and align assessments by grade, level and subject.* Under the leadership of the Design Team, the district's leading content and assessment specialists should be convened to ensure that the curriculum and designated assessments are aligned with district standards. Further, they should ensure that all teachers understand and are trained to use those standards when setting objectives, delivering instruction and monitoring progress.

IMPACT

Teachers have demonstrated their willingness to field-test a new approach to linking student achievement to teacher compensation. This approach would be greatly strengthened by aligning the assessments with their intended usages. Further, this will increase the levels of fairness within the pilot.

Issue Four: *Professional Development*

OVERVIEW

In both interviews and surveys, teachers and principals are indicating their needs for increased professional development. They are appreciative of the responsiveness of the Design Team, while demonstrating the need for expanded levels of support from the district.

Teacher requests for assistance reflect the importance of providing professional development that is sustained, on site and based on the differentiated needs of the schools. There is a high level of teacher interest in professional development. This extends to such areas as setting objectives, understanding student achievement data, and developing and implementing new teaching strategies. Further, principals have identified a salient need for training in classroom observations and monitoring. This is an opportune time for a focused district response.

RECOMMENDED ACTION

- *Provide expanded professional development for pilot teachers and principals.* This should focus on the teaching and learning content issues identified through the pilot. There is a particular need to ensure that teachers have the tools to align the objectives with district standards and measurements, understand the needs and best practices to help low achieving students, and provide differentiated instruction to students, as needed. The district should further embrace the diversity of the students by providing the training and resources that will help teachers to succeed with students who require more or different kinds of instruction.
- *Establish the district leaders for professional development.* This includes defining the structure that will oversee and coordinate professional development efforts district-wide. This is needed for the district as a whole, as well as for departments and by content areas.
- *Conduct a professional development audit.* Particularly in a period of limited resources, it is essential to know the scope, financial costs, resource allocations, sources and effectiveness of the professional development currently being provided within the district, so that district resources can be used to the greatest effect.
- *Prepare for the post-pilot transition.* To date, the Design Team has fulfilled multiple leadership roles, from identifying needs to ensuring the delivery of training to monitoring quality. Should the pilot move to scale, these functions will need to be filled by central departments operating with the same sense of urgency and mission as the Design Team.

IMPACT

The twin pillars of reform are accountability and support. Denver teachers have embraced the responsibilities related to accountability. This commitment needs to be reinforced by an intensive delivery of professional development.

Issue Five: *The Pilot and a New Salary Structure*

OVERVIEW

The pilot has significant potential to provide learnings that will be valuable to the district when developing a new salary structure. However, there are constraints which affect the potential benefits of the pilot in this regard. First, the pilot is field-testing an approach that is based on providing relatively small bonuses on top of the existing salary structure. Yet the sponsoring parties are not seeking to implement a bonus system in the long-term. Second, in terms of the fairness, precision and accuracy of measures, there are constraints on the ability to measure a teacher's contribution to student achievement when approached in one year increments. In the ensuing years of the pilot, these constraints will be tested to ascertain trends and develop resolutions to the problem, but this is already an emerging area of concern. Third, and most importantly, the pilot is being approached more as a driver of a new compensation system, rather than as an essential component of such a system.

The sponsoring parties and the participants are interested in maximizing the benefits of the pilot. In terms of the future compensation system, the pilot's statement of purpose provides a vehicle for doing just that. The statement indicates that the purpose is to develop a teacher compensation system that is based, in part, on student achievement. Compensation is generally perceived as occurring in annual increments. However, in the context of the experimental nature of a pilot, the term—"in part"—opens up opportunities for the district to view compensation from an innovative perspective.

RECOMMENDED ACTION

- *Reposition the pilot in relation to a new teacher salary structure.* The Joint Task Force is charged with researching promising practices in teacher compensation from around the nation. Task force findings will be channeled into the collective bargaining negotiations. As part of this process, and in response to the fairness issues cited above and throughout this report, the sponsoring parties might consider the possibility of basing *part* of the compensation on pay for performance—as determined by the performance of teachers over a multi-year period. This would maintain the intent of pay for performance, while addressing issues of fairness and avoiding unnecessary controversy.

IMPACT

The pilot focuses on the goal of linking student achievement and teacher compensation. In so doing, Denver need not be burdened by the unrealistic goal of developing the perfect system for learning and finance. Rather, the district needs to continue to build on the pilot's accomplishments thus far and make viable improvements to the current method of compensation.

Issue Six: *The Approaches*

OVERVIEW

The pilot is field-testing three approaches, using norm-referenced tests, criterion-referenced tests and demonstrations of teacher acquisition of skills and knowledge. These were initially seen as two output approaches and one input approach. However, all of the approaches now have linkages to student achievement measures. Moreover, all three approaches require providing professional development to teachers and site administrators so that they can establish better objectives, understand student achievement more thoroughly and convert learnings into improved pedagogy.

At the mid-point of the pilot, the three approaches are showing varying results. However, a single approach has not emerged as superior in terms of increasing student achievement. Further, teachers in all of the approaches question the fairness of judging performance based on a single academic measure. Together, these findings present an opportunity to further refine the pilot.

RECOMMENDED ACTION

- *Begin to integrate the three approaches.* This is a longer-term effort. It would involve assessing classroom results with multiple measures and providing teachers with the necessary professional development. This would result in a more focused and fair initiative.
- *Determine the method and timeframe for integrating the three approaches.* Under the leadership of the Chief Academic Officer and the Design Team, convene site and central level practitioners for this purpose.

IMPACT

The district is committed to becoming a learning organization. By integrating the three approaches, the district will demonstrate the application of several key findings. Further, the result will be a pay for performance mechanism that is fairer, more manageable and easier to support with professional development.

Issue Seven: *Pilot Requirements and Financial Forecasts*

OVERVIEW

There is a range of costs connected to implementing a pilot and making systemic changes. These costs take two forms. First, there are costs which are financial in nature. These result from new fiscal outlays such as salaries, equipment and additional staffing. Second, there are costs related to changing practices. These are non-financial in nature and include the institutional costs of reordering district priorities, functioning with higher levels of interdepartmental coordination and reallocating existing funds.

Through years of contract negotiations, the Board of Education and the Association are highly familiar with the financial costs of change. By putting key organizational issues on center stage, the pilot is also bringing attention to issues requiring new priorities and new commitments. With a pilot as far reaching as Pay for Performance, it is essential to have short- and long-term projections of the financial and non-financial costs of implementation.

RECOMMENDED ACTION

- *Project the costs of implementation.* A two-year projection covering the balance of the pilot and a five-year projection covering the costs of bringing the pilot to scale will help the district identify critical and immediate needs and plan appropriately for the potential changes that might come from bringing a new compensation system to scale.
- *Ensure the inclusion of both direct and indirect costs.* These should include the related costs at the school sites, and such units and departments as Assessment and Testing, Curriculum and Instruction, Professional Development, Technology Services, Planning, Human Resources and related systems.

IMPACT

In 2003, the district will be making major policy decisions regarding the future of Pay for Performance in the district. The short- and long-term projections of costs will inform policymaking, provide information essential for making tough managerial choices and red flag issues well in advance of contract negotiations. These are all benefits.

B. Summary

The Board of Education and the Association have moved the Denver Public Schools to the forefront of educational reform in the United States. They have established a pilot that is unique. Moreover, the parties have demonstrated their joint emphasis on results by committing to study the impact of the pilot.

The challenges for the Board of Education and the Association are to address the learnings from pilot activity to date, and to begin planning for the longer term even as the pilot continues to evolve over the next two years. The Denver educational community has shown a willingness to look forward and take critically needed, strategic risks on behalf of students. Making the best use of this landmark pilot will involve all parties in expanding the efforts to align curriculum, instruction and assessment to strengthen classroom instruction, and to evaluate classroom results as a part of the teacher compensation system. Denver is forging a new pathway to results.

Notes

Endnotes

Chapter II

1 United States National Commission on Excellence in Education. *A Nation At Risk: The Imperative for Educational Reform: a report to the Nation and the Secretary of Education, United States Department of Education/by the National Commission on Excellence in Education.* The Commission: [Supt. Of Docs. U.S.G. P.O.] 1983.

2 Wellford Wilms and Richard Chapleau, Education Week, Nov. 3, 1999, "The Illusion of Paying Teachers for Student Performance—Some Lessons from History"

3 Alan Odden, ERIC Digest 142, Nov., 2000

4 Wellford Wilms and Richard Chapleau, Education Week, Nov. 3, 1999, "The Illusion of Paying Teachers for Student Performance—Some Lessons from History"

5 Alan Odden, ERIC Digest 142, Nov., 2000

6 Jeff Archer, Education Week, 2/7/01, "Business Seeks Teacher Renaissance"

7 Jeff Archer, Education Week, 6/21/2000

8 Education Week, 9/19/01

9 Lynn Olson, Education Week, 10/13/99, " Pay-Performance Link in Salaries Gains Momentum"

10 Ann Bradley, Education Week, 2/25/98

11 The Washington Post, 6/18/00, pg. A20

12 DPS Public Information Office, Nov. 1, 1999 Background Paper

13 1999-2000 Agreement Between Denver Classroom Teachers Association and School District No. 1, in the City and County of Denver and State of Colorado, Effective Sept. 1, 1999 to Aug. 31, 2000 (Hereafter referred to as "Agreement")

14 Previously cited Agreement, Appendix E, Currently Appendix 01-01(1), III-A-3

15 Previously cited Agreement, Appendix E, Currently Appendix 01-04(1), III-A-2a

Chapter III

1 Merriam, S. B. *Case Study Research in Higher Education.* San Francisco: Jossey-Bass, 1988.

Stake, R.M. The Case Study Method of Social Inquiry. In G. F. Madaus, M. Scriven, and D. L. Stufflebeam (Eds), *Evaluation Models.* Boston: Kluwer-Hijhoff Publishing, pp. 279-2861991.

Yin, R. K. *Case Study Research: Design and Methods.* Beverly Hills, CA: Sage, 1984.

Chapter IV

1 The Design Team Project Plan, Component Two: Objectives (January 2000).

2 Jane Stallings *et al, Elementary School Journal,* (v86, n5 (May 1986), "Effects of Instruction Based on the Madeline Hunter Model for Students' Achievement: Findings from a Follow-Through Project," 571-87. See also Pat Wolfe, *Educational Leadership,* v56, n3 (Nov 1998), "Revisiting Effective Teaching," 61-64.

3 See links for teacher lesson planning at the following: http:/www.askeric.org/Virtual/lessons/

4 School plans and control school objectives were not available for 1999-2000.

5 Grant Wiggins and Jay McTighe, Understanding By Design. (Association for Supervision and Curriculum Development: 1998), p. 9.

6 See these standards at http:/www.mcrel.org/.

7 Slight discrepancies in numbers may result from data provided at different points in time or from incomplete data.

Chapter VI

1 Bryk, A. S., & Raudenbush, S. W, *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage, 1992.

2 Singer, J., "Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models," *Journal of Educational and Behavioral Statistics, 24* (4), pp. 323-355, Winter, 1998.

Chapter IX

1 *Education Week,* May 23, 2001; "Study Questions Reliability of Single Year Test-Score Gains", Lynn Olson

2 National Research Council, High Stakes Testing: Testing for Tracking, Promotion and Graduation, National Academy Press, Washington D.C., 1999, p. 3.

3 Ibid, p. 6.

Notes

Notes

ctac COMMUNITY TRAINING
AND ASSISTANCE CENTER

30 WINTER STREET • BOSTON, MA 02108
TEL: 617.423.1444 • E-MAIL: ctac@ctacusa.com