Catalyst for Change

Pay for Performance in Denver
Final Report

January 2004

COMMUNITY TRAINING
AND ASSISTANCE CENTER

ABOUT CTAC:

The Community Training and Assistance Center is a national not-for-profit organization with a demonstrated 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. Since 1979, CTAC has provided assistance to hundreds of community-based organizations, coalitions and public institutions in the United States and several other countries.

CTAC'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.

Catalyst for Change

Pay for Performance in Denver
Final Report

January 2004

COMMUNITY TRAINING
AND ASSISTANCE CENTER

Acknowledgements

The Community Training and Assistance Center wishes to thank the sponsors of this research study, the Denver Public Schools Board of Education and the Denver Classroom Teachers Association. Their commitment to bringing quality and accountability to public education is exemplary. During the four years of the pilot, they have been able to maintain a critical focus on the development and implementation of Pay for Performance.

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 Les Woodward, former President Elaine Gantz Berman, Lucia Guzman, Bruce Hoyt, Michelle Moss, Kevin Patterson, Theresa Peña, and previous board members Sue Edwards, Sharon Macdonald, James Mejia and Bennie Milliner. The leadership of the Association has also been essential, including President Rebecca Wissink, former President Andrea Giunta and Executive Director Bruce Dickinson. Thank you for being willing to share this professional challenge with CTAC; we have appreciated your support and confidence.

The CTAC research study provides a solid foundation for the ideas and desires of those who wish to create a new design for a public school system. We wish to thank those individuals who, during these past four years, have contributed so much to this effort. 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.

The coordination and implementation of the many facets of 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 Rebecca Wissink, and assistant Ginger Doyle. Their drive, passion and sense of urgency are inspiring.

We would like to extend a special thank you to the unsung heroes of the pilot and study—the pilot school teachers, administrators and parents. They have made 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 Superintendents Richard Allen, Wayne Eckerling and Andre Pettigrew, Chief Academic Officer Sally Mentor Hay, and former administrators Beth Celva, Craig Cook, Larry Lindauer and Dianne Lefly. Key contributions have also been made by James McIntosh and Richard Machado of Assessment and Testing; Ethan Hemming of Educational Services; Joyce Fell, Peggy Flynn and Billy Gould of Human Resources; and Director Aaron Gray and Flor Amaro of Community Relations. They have all been generous with their time and attention.

There were two groups which have assumed additional responsibility for the progress of the pilot. They are the pilot Leadership Team and the Joint Task Force on Teacher Compensation. The participating individuals have dedicated an inordinate amount of their professional and personal time to understand and facilitate important elements of this project.

CTAC is grateful for all those listed above, and the many other people who helped to make this study a reality.

This study has been made possible through the generous support of Rose Community Foundation, The Broad Foundation, The Daniels Fund, The Sturm Family Foundation, Jay and Rose Phillips Family Foundation, The Denver Foundation, Donnell-Kay Foundation, and The Piton Foundation.

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

Principal Study Authors
William J. Slotnik
Maribeth D. Smith

Contributing Authors
Roberta J. Glass
Barbara J. Helms, Ph.D.

Field Operations Director
Donald W. Ingwerson, Ph.D.

Team Members
Denise A. Bell
Lee Bray
Peggie L. Brown
Robin C. Burr, Ph.D.
Mary Ann Crayton
William M. Eglinton
Donald B. Gratz, Ph.D.
Mimi Howard
William C. Lannon
Juan J. Maya
Candy Miller
Marcia J. Plumleigh, Ph.D.
Lynn J. Stinnette
Martha Swartz
Julia D. Thomé
Barbara Zeno

Statistical Measurement Task Force
Robert H. Meyer, Ph.D.
John B. Willett, Ph.D.

Statistical Review
Kathleen Calore

Contents

Executive Summary

The Denver Public Schools (DPS) and the Denver Classroom Teachers Association (DCTA) jointly sponsored the Pay for Performance (PFP) pilot. This four-year pilot, conducted during the 1999-2003 school years, focused on developing a direct link between student achievement and teacher compensation. The sponsoring parties committed to implementing the pilot and to studying the results of this initiative. *Catalyst for Change* is the final report of the results of this innovative collaboration on behalf of students and teachers.

The centerpiece of Pay for Performance in Denver has been the teacher objectives at the 16 pilot schools. Teachers developed two annual objectives based on student achievement that required the approval of the principal. Teachers received additional compensation if they met their objectives. The entire pilot was stewarded by a four-person Design Team that included district and union appointees. The pilot involved 13% of Denver's schools. These schools represented a cross section of the district's client and service base. In microcosm, the pilot schools presented the challenges of the broader district. Their experiences provided a foundation for a study of pilot impact.

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 the technical assistance necessary to help assure a pilot of quality and integrity. The study's mid-point report, *Pathway to Results: Pay for Performance in Denver* was presented in December 2001. *Catalyst for Change* is the summative report.

The study has four core components. First, it examines the impact of the pilot on student achievement based on two independent assessments. Second, it examines teacher objectives: their substance, quality and relationship to student achievement. Third, the study considers the effect of a range of student, teacher, and school factors on the results of the pilot. Finally, it identifies the broader institutional factors that have affected implementation.

The data for this four-year study has been extensive. CTAC conducted surveys and examined responses from 2,870 teachers, parents, administrators and others involved in the pilot; conducted and reviewed 615 interviews; and organized and analyzed tens of thousands of student records plus teacher and demographic data for the pilot and control schools. The study also involved the careful review of artifact and documentary data and the detailed examination of 4,012 teacher objectives. Human resource records and school documentation were linked to school, teacher, and student information to create a comprehensive database. Finally, the study included hundreds of hours of observations, from classrooms to boardrooms, which contributed a strong sense of the context and the everyday work of the pilot.

The nature of a pilot is experimentation. Denver conducted the pilot in order to examine the effectiveness and impact of a new direction. By building around the objective setting process, the pilot design was straightforward and honored both teacher judgment and perspective. The implementation of the pilot, though, was necessarily more complex. As teachers were learning about developing and meeting measurable annual objectives, the schools and the district were learning about the necessary alignment of the curriculum, assessment, student data, human resources and other parts of the larger system with Pay for Performance. The alignment and strengthening of these functions proved complicated and challenging.

The pilot has demonstrated that the focus on student achievement and a teacher's contribution to such achievement can be a major trigger for change—if the initiative also addresses the district factors that shape the schools. The findings show promising results in many of the areas studied. However, the pilot's implementation also revealed areas of practice and policy that will need to be strengthened further in order to continue improving student achievement as the key elements of Pay for Performance are taken to scale in the district.

Highlighted below are CTAC's analyses, findings and recommendations. All of the recommendations are crafted to meet the standard of increasing the manageability, fairness and sustainability of Pay for Performance. The report also examines the national implications of Denver's Pay for Performance pilot. 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

- At all three academic levels—elementary, middle, and high school—higher mean student achievement in the pilot schools is positively associated with the highest quality objectives. Students whose teachers had excellent objectives, based on a four-level rubric developed by CTAC, achieved higher mean scores than students whose teachers' objectives were scored lower on the rubric. This holds true on most tests of the *Iowa Test of Basic Skills (ITBS)* and the *Colorado Student Assessment Program (CSAP).*

  At the elementary school level, students of teachers with excellent objectives (rubric level 4) had significantly higher mean normal curve equivalent (NCE) scores on the *ITBS* Reading, *ITBS* Language, and *CSAP* Math tests than students of teachers with lower rubric levels.

  At the middle school level, students of teachers with level 4 objectives had significantly higher mean scores than students of teachers with level 3 and level 2 objectives on the *CSAP* Math test.

  At the high school level, there were significantly higher mean scores on the *ITBS* Reading and *CSAP* Writing tests at Manual High School and on the *CSAP* Writing test at Thomas Jefferson High School for students whose teachers had level 4 objectives.

  Six other tests (one elementary, three middle school, and two high school level) show a positive relationship between highest quality objectives and higher achievement level that is not statistically significant.

- Pilot students were compared to control students, estimating the change in mean NCE scores over time (e.g., from the baseline year through the end of the pilot) on three tests of the *ITBS* and three tests of the *CSAP.* Two-stage hierarchical linear modeling (HLM) was used to account for correlation between observations and to adjust for differences in school and student characteristics. The use of NCEs makes it possible to determine whether students are performing better than expected based on previous scores—attaining more than an expected year of growth—or not.

  The effect of the pilot varies by level of school:

  The pilot elementary school mean NCE scores declined on the *ITBS* Math, *CSAP* Reading and *CSAP* Math tests. The control school mean NCE scores declined on the *ITBS* Math test and increased on the *CSAP* Writing test. These results were statistically significant. The performance of the pilot students was lower than the controls on the tests except on the *ITBS* Language.

  The pilot middle school students performed significantly higher than the controls on the *ITBS* Reading, *CSAP* Writing, and *CSAP* Math tests. The average NCE scores of middle school pilot students increased significantly over time (i.e., attained more than a year's expected growth) on the *ITBS* Reading, *CSAP* Writing, and *CSAP* Math tests. The controls experienced statistically significant declines in mean NCE scores on the three *ITBS* tests over the course of the pilot and statistically significant increases on the three *CSAP* tests. On the *ITBS* Language test, both the pilots and controls decreased by comparable amounts.

The high school pilots and controls experienced statistically significant increases in mean NCE scores over time on most tests. Manual High School students had significantly higher increases than the control school students on the *ITBS* Language and Math tests. Thomas Jefferson High School students performed significantly higher than the control school students on the *ITBS* Language, *ITBS* Math and the *CSAP* Reading tests and significantly lower on the *ITBS* Reading test.

- Meeting two objectives is positively associated with higher mean achievement scores.

  At the elementary schools, the students of teachers who met two objectives had significantly higher mean NCE scores on all six tests than students of teachers who met one objective.

  At the middle schools, meeting one or two objectives was associated with significantly higher mean NCE scores than meeting no objectives on the *ITBS* Reading and Language tests. Similar positive relationships were seen on the *ITBS* Math and *CSAP* Math tests but they are not statistically significant.

  At the high schools, the students of teachers who met two objectives (at both Manual and Thomas Jefferson) had significantly higher mean NCE scores than students of teachers who met one objective or no objectives on the *ITBS* Reading test. Similar positive relationships are found on the *ITBS* Math, *CSAP* Writing, and *CSAP* Math tests at Manual High School and on *ITBS* Math, *CSAP* Reading, and *CSAP* Math at Thomas Jefferson High School but they are not statistically significant.

- Student achievement rises as length of teacher participation in the pilot rises.

  Elementary students whose teacher had been in the pilot for two, three, and four years had mean *ITBS* Reading scores 0.8, 1.3, and 2.2 NCEs higher than students of one-year teachers. Elementary students of four-year teachers also had mean *ITBS* Math scores significantly higher than those of one-year teachers.

  Middle school students of two-year teachers scored 2 NCEs higher on average and students of three-year teachers scored 3.2 NCEs higher than students of one-year teachers.

  High school students of two-year pilot teachers scored higher on *ITBS* Reading, *ITBS* Math, and *CSAP* Reading at both pilot high schools, but the difference was only statistically significant for the Thomas Jefferson *ITBS* Reading exam.

- The pilot has been the catalyst for developing a fundamentally new compensation plan for teachers in Denver which is based, in part, on student achievement.

Impact of Objectives

- The percent of teachers who developed objectives that were rated at the two highest levels of the rubric increased steadily over the course of the pilot. The particular improvement in the objectives in the final year of the pilot is largely attributable to greater attention to learning content in the objectives. By the close of the pilot, 28% of the objectives were at level four (excellent) and 44% were at level three (adequate).

- There is a significant increase in the quality of the objectives as the number of years a classroom teacher participated in the pilot increases.

- Teachers met their objectives at a high rate. The data collected by the district over the four years of the pilot show that from 89% to 93% of the teachers met one or more objectives and were awarded additional compensation.

- As teachers gained more years of experience in the pilot, their ability to meet their objectives also increased significantly. One-year pilot participants met 89% of their objectives. The success rate rose to 98% for four-year participants.
- Educational background and years of experience in the Denver Public Schools are related to whether classroom teachers met their objectives.

  Certified teachers met 92% of their objectives, while teachers participating in the Teacher-in-Residence alternative certification program met 83% of their objectives.

  First year teachers met 86% of their objectives, while teachers with two or more years of experience met 92% of their objectives.

  Teachers with 15 or more years of experience in the Denver Public Schools met their objectives at lower rates (85%) than teachers with fewer than four years (95%), four to 10 years (90%), or 11 to 14 years (95%) of experience.
- There are similarities between pilot school teacher objectives and control school teacher goals. However, the control school teachers make less use of baseline data. Some of the similarities are attributable to the pilot's objective setting protocol being in full or partial use in nearly one-third (32%) of the control school goals reviewed in 2002-2003.
- The objectives and their learning content are not included as part of the strategies in the school improvement plans.

Perceptions of Participants and Other Parties

- The pilot has significantly increased the school and district focus on student achievement. This focus has increased with each succeeding year of pilot implementation.
- Teachers indicate that they have greater access to student achievement data and that they use the data more effectively, particularly baseline data, to establish growth expectations, to focus earlier on students who may need more assistance and to monitor progress.
- Most pilot teachers do not attribute changes in their core classroom instructional practices to the pilot. Teachers indicate that they did not receive a mandate to make such changes.
- Most teachers feel that cooperation among teachers has improved or stayed the same at the pilot schools.
- Pilot teachers are less fearful of pay for performance than control school teachers. By the end of the pilot, pilot participants were more likely to offer suggestions for improvement than to indicate that pay for performance was not viable.
- Pilot teachers continued throughout the pilot to raise issues of fairness and trust in the objective setting and review process. However, they believe that it is possible to set fair objectives.
- The quality of interaction between the principals and teachers is pivotal to the implementation of Pay for Performance at the school sites. Teachers in the pilot schools believe that there are inconsistencies from school to school in how principals review and analyze progress on teacher objectives. Principals indicate that there is a lack of clarity regarding their role and authority and a need for targeted professional development.
- Parents indicate that a teacher's contribution to student achievement should be rewarded in financial terms.

- The vast majority of parents (94%) and teachers (93%) feel that more than one measure of student achievement should be used to determine teacher performance.
- Participants value the training they received, but express a need for more professional development based on the specific student achievement levels in the individual schools and classrooms and the instructional challenges of meeting objectives.

Institutional Factors

- The DPS/DCTA collaboration on behalf of student achievement has been significant. This collaboration has been pivotal to the development of Pay for Performance despite changes in district leadership and structure.
- Pay for Performance has enabled issues which have adversely affected district progress, sometimes for many years, to be put on center stage. Operating in a climate protected by external supporters and internal reformers, the pilot provided a vehicle for problems to be discussed, analyzed and acted upon. These actions have helped the district to develop an increased capacity to make mid-course corrections.
- Teachers and principals were provided with multiple opportunities through the study to influence the course of the pilot. For many, this was a marked and positive departure from past district practice.
- The Design Team contributed significantly to the progress of the pilot.
- District support systems were seriously challenged by the implementation of Pay for Performance. Many opportunities for change were identified and district action resulted. Challenges of organizational alignment still lie ahead for the district.
- The turnover in leadership positions during the course of the pilot, particularly at the level of the pilot school principals and the superintendency, contributed to some of the concerns related to trust and institutional priority that have affected the implementation of the pilot.
- The lack of an agreed-upon and aligned portfolio of district assessments for measuring student achievement meant that 166 identifiable assessments were used to measure progress in meeting objectives, and 256 teachers used generally referenced measurements, in the last year of the pilot.
- The task of linking student achievement results to specific teachers has proven more challenging than originally anticipated by the district. As pilot efforts go to a broader scale of implementation in the district, this type of data capacity will be greatly needed.
- Several factors, including the state and national high stakes testing environment and the district's experiences with pay for performance for administrators, adversely affected the climate for implementing the pilot.

B. Recommendations

Issue One: Alignment

Since the purpose of the district's major initiatives is to increase student achievement, the organization will benefit from continuing to align its initiatives around that goal in a clear and purposeful manner.

Recommendations include:

- *Bring the objective setting to scale with instructional support.* Crafting objectives is a key initial step in planning and delivering instruction. It is not merely an exercise in writing. It will be important to align instructional support to assist teachers to meet the specific targets in their objectives.

- *Strengthen the linkage between classroom objectives, school improvement plans and district standards and goals.* To the extent that objectives, school plans, and standards and goals are mutually reinforcing, the implementation of all three will be strengthened and there will be greater clarity of purpose.
- *Increase the connection between student information systems and human resources systems.* Building on the district's progress in supporting the pilot schools, this recommendation focuses on establishing greater district-wide linkages among the data systems for student assessment, planning, and human resources. These linkages are critical for any effort that seeks to examine the contribution of a teacher to student achievement.
- *Project the costs of changing internal practices and requirements.* There are direct financial costs to implementing Pay for Performance systemwide. To maintain teacher commitment, the momentum of the pilot needs to continue under a range of financial circumstances.

Issue Two: Assessment

A portfolio and appropriate usage of high quality assessments are the marks of a district that is aligned and accountable for its students. Using student assessments for a new purpose—compensation—requires greater assessment and data capacity, as well as a skillfully supervised and consistently administered effort at the school sites so that all students have regular assessments.

Recommendations include:

- *Expand the district's assessment strategy.* The existing assessment plan should become a component of a more comprehensive assessment strategy that includes aligning fair and valid assessments with the district's standards and curricula, ensuring practitioner input into the selection and use of assessments, and establishing a clear direction about who is to be assessed.
- *Define which assessments can be used for objective setting and compensation purposes.* There continues to be a need for a rational level of prescriptive direction regarding which assessments can be used as part of any new system that involves pay for performance. A pay for performance system or companion educational initiative that has too many allowable assessments will be unmanageable, will cause discord and will fail to promote valid increases in student achievement.
- *Make the use of multiple measures a developmental priority.* For four years, Denver teachers and site administrators have been raising questions about the fairness and accuracy of single measures. The charge is for the district to develop a means to link several assessments together to more meaningfully identify student progress.
- *Increase the district capacity to disaggregate and analyze student achievement data.* Regular analyses of these data strengthen decisions about delivering classroom instruction, developing school improvement plans, and managing strategically at the district level and, in the era of No Child Left Behind, it will provide communities with high quality information about its schools.
- *Convene select urban districts to analyze and take action on problems in assessments.* As a result of the Pay for Performance pilot, Denver is positioned as a national leader in the area of tying teacher compensation, in part, to student achievement. The ensuing challenges that Denver faces are shared in common by other districts. Denver should convene a small number of urban districts and assessment specialists to guide test developers to link their efforts more directly to the growing needs of urban districts.

Issue Three: Professional Development

Virtuosity in teaching is the goal of professional development for teachers. Initiatives often assume that teachers will embrace the concept of the reform and change their practices when, in fact, they may not know about new practices that would be appropriate. Both educational research and the pilot outcomes indicate that there is a profound connection between objectives based on learning content, a teacher's subject matter knowledge, specific teaching practices, and student achievement.

Recommendations include:

- *Establish district standards for professional development.* Establishing quality standards for professional development is a natural and necessary complement to instructional priorities. They should be tied to the Colorado Teacher Standards, research about best teaching practices, the district's curriculum standards, and the assessment strategy described above. This work will result in a roadmap for providing professional development services and ensuring quality control.
- *Predicate professional development on student achievement.* The priorities for professional development need to be based on continuous reviews of student achievement results by school staffs. Such a review identifies schoolwide, classroom and individual student instructional needs and instructional areas which need to be updated or improved. This, in turn, may reveal areas in which school staff or the community may need assistance in meeting these needs.
- *Create opportunities for teachers and principals to shape professional development.* One of the key findings from the pilot is that the ability of site practitioners to influence implementation contributes to improvements in the overall effort. This kind of involvement increases the prospects of professional development to effectively target teacher needs, school priorities and district goals.

Issue Four: Leadership

Many reforms fail for lack of sustained leadership. The Board of Education and the Association demonstrated leadership as they joined to create the pilot. The Design Team has provided creative leadership in advancing the pilot through uncharted pathways. As the effort moves forward to institutionalize the critical elements of the pilot into district practice, quality leadership will be essential to shape and guide the reform through its next steps.

Recommendations include:

- *Broaden the collaboration on behalf of student achievement.* The pilot is the result of an unprecedented collaboration between the Board of Education and the Association. This collaboration has been substantive and effective. It should be extended to other parts of district educational operations, regardless of the outcome of the Association and Board votes on a new compensation plan.
- *Continue to place problems on center stage.* A central factor contributing to the accomplishments of the pilot has been the ability to place critical issues that affect the district on center stage. The district will benefit by continuing and extending this function.
- *Create a Principals Leadership and Achievement Institute.* All principals need to understand deeply how learning occurs and how it is nourished, measured and supported. They need ongoing, sustained opportunities to identify salient site issues, analyze trends in student achievement data, reflect on emerging issues, develop their skill in observing classrooms and providing support to teachers, and build the knowledge to work effectively with diverse students and families. Building these capacities will complement the current district plans to prepare principals to carry out targeted educational initiatives.

- *Prepare for the post-pilot and post-vote transition.* The pilot benefited greatly from having a special internal implementation team with the commitment and sense of urgency that is essential to create change. As the learnings and practices from the pilot are implemented district-wide, it will be essential to institutionalize the qualities that the Design Team brought to the implementation of the pilot.

C. Summary

A major initiative that focuses on student achievement—while concurrently exploring changes in the teacher compensation system—goes to the heart of the district mission and structure. As such, the Pay for Performance pilot and study were significant undertakings.

Denver introduced Pay for Performance as a new element in a large urban district. The pilot has been a catalyst for changing the district so that it could become focused on student achievement in a more coordinated and consolidated way. A key part of Denver's story is how a pilot, with key internal and external supporters, engendered positive change in a larger institution. Many of the changes have been systemic—changing how the system thinks and behaves. They are, though, works in progress. Challenges of organizational alignment remain distinct.

As in many large urban districts, Denver experienced leadership transitions over the four years of the pilot. However, the Board of Education and the Association stayed the course. As a result, the pilot achieved a substantial degree of reach into the system. By so doing, Denver has contributed to its own systemic improvement efforts as well as to those of other districts who may want to go down this path.

The issue of aligning a district in support of a pay for performance system cuts to the very essence of how—and to what extent—a school district is functioning in support of student learning. The changes required to identify, strengthen and reward individual student growth and individual teacher contributions under pay for performance have the added effect of stimulating other parts of the school system to improve the quality of support and service. The result is a *catalyst for change* that benefits all students and teachers.

CHAPTER I

Overview

A. Background and Charge

In September 1999, the Denver Public Schools (DPS) and the Denver Classroom Teachers Association (DCTA) embarked on what would become a four-year pilot and study of Pay for Performance in 16 schools in the district. An initiative in teacher improvement and accountability, the pilot was established to develop a link between teacher compensation and student achievement through a design that came out of the negotiations process and was captured in the negotiated agreement between the district and its teachers.

Efforts to institute performance or incentive pay for teachers have a record of unsuccessful implementations and have characteristically been anathema to teacher organizations and "folly" to many teacher researchers.[1] For this reason many eyes have been on Denver, as the district and the Association collaborated to design and implement a pilot that would overcome some of the well-documented objections to pay for performance in education and, additionally, lead to improvements in student achievement.

Coming at a time of increased accountability measures coupled with a scarcity of qualified teachers, the pilot in Denver recognized that teachers are the critical link to children achieving high standards and that compensation schedules should reflect this fact. Aiming at more rigorous standards for students requires teachers who are capable of transmitting deeper knowledge and greater skills to their students. According to a recent national analysis:

> "State education leaders recognize that teaching, perhaps more than any other element of a child's education that occurs at school, is critical to achieving high standards. To bolster the professionalism of the teaching field, meaningful salary increases must be tied to improvements in teacher performance."[2]

At the core of the Pay for Performance pilot is a process whereby teachers set two classroom-specific objectives with the approval of the building principal and then present evidence of attainment to the principal for verification at the

end of the year. If the evidence substantiated that the teacher had met one or both of his/her objectives, then a performance bonus per objective was awarded to the teacher. Other significant features of the negotiated agreement between the district and the Association were (1) the authorization of the Design Team as the stewards of the pilot and (2) the commission of a comprehensive research study to explore the impact of the pilot and the effect of a range of contributing factors on the outcome of the pilot. A third important negotiated feature was introduced later in a separate memorandum of understanding—the establishment of the Joint Task Force on Teacher Salary (later the Joint Task Force on Teacher Compensation), formed for the purpose of designing and recommending for adoption a new compensation plan that would be based, in part, on student achievement.

The charge of the four-member Design Team was two-pronged: to develop the pilot as a study of the "feasibility of linking student achievement to teacher compensation" and to evaluate the "capacity of the school system to implement such a program successfully should it be adopted."[3] To help meet this charge, the Community Training and Assistance Center (CTAC) was selected to conduct a study of the impact of the pilot and to provide technical assistance that would help build district capacity to implement a pilot of quality and integrity. CTAC is a national non-profit organization, based in Boston, which has been a leading provider of technical assistance and research services to community-based organizations, coalitions, and public institutions in the United States and several other countries for twenty-five years. In this role, CTAC has worked extensively with school districts that are attempting to improve student achievement, community involvement, and overall school and district performance and accountability.

School participation in Pay for Performance was voluntary, based on faculty votes. During the first year 12 elementary schools entered the pilot. In the second year a middle school entered the pilot. By the close of the pilot in June 2003, 16 schools were participating in Pay for Performance.

Originally the negotiated agreement identified three approaches to be compared in the pilot. Schools entered the pilot designated as one of the following: (1) an Approach One school, which measured student progress on a norm-referenced test; (2) an Approach Two school, which measured student progress on a criterion-referenced test or teacher-created measures; or (3) an Approach Three school, which focused on the teachers' acquisition of skills and knowledge. These were seen as two output approaches and one input approach. At the mid-point of the pilot, the three approaches were integrated into one because all approaches were linked to student outcome measures and all teachers required professional development opportunities. In addition, no significant difference among the approaches had emerged in the first two years of data.

B. Areas of Inquiry

The study of Pay for Performance, as conducted by CTAC, examines four overarching and interacting areas of the pilot, which collectively focus on results and the key factors that may contribute to these results.

Impact on Student Achievement

The focus of the pilot, and concomitantly of the study, is student achievement. Individual student growth (from spring to spring) is the basic unit of measurement in the study. The study examines changes in student achievement that have occurred in the participating schools in comparison to those in the designated control schools, as well as how student achievement gains correlate to the quality and attainment of teacher objectives. For this purpose, the study uses student achievement data from the *Iowa Test of Basic Skills (ITBS)* and the *Colorado Student Assessment Program (CSAP)*. The *ITBS* is a national norm-referenced assessment for grades 2-11; the *CSAP* is the Colorado standards-based assessment that has been phased in by grade level during the course of the pilot.

Impact of the Objectives

The spotlight of the pilot has been on the teacher-developed objectives through which additional compensation may be earned. The study examines the quality and rigor of the objectives, their impact on student achievement, and the perceptions of pilot participants about the nature and effect of

objectives in the school setting. Additionally, objectives are considered from the perspective of whether the teacher met them based on their own measurements and from the perspective of student growth in each teacher's classroom at the elementary school level.

School, Teacher and Student Factors

The study examines school, teacher, and student factors for their potential contribution to the achievement of students and the outcomes of the pilot. The schools participating in the pilot serve different populations, therefore it was necessary to control for student and school characteristics. The study also explored the relationship between achievement outcomes and teacher characteristics. Additionally, factors such as school plans, teacher experience, and leadership quality and stability are examples of potential influences on the results of the pilot.

Broader Institutional Factors

The study examines a range of institutional factors that have influenced the outcomes of the pilot and from which important lessons can be derived. For example, the availability of adequate and reliable measures for teacher use, as well as alignment between standards, assessments, and professional development, and the availability and access to student data for teachers are significant systemic factors that affect a pilot of this nature.

C. Data Components

The study of the impact of the pilot is based on several primary sets of data, collected in each of the four years of the study, which have been subjected to several layers of analysis. A brief description of data sources is provided below and sources are referenced throughout the text of the report. Data components for the study include:

Comprehensive Surveys

Confidential surveys of participants, including teachers, administrators and parents at both the pilot and control schools, elicited the perspectives of a range of stakeholders on the status of the pilot, including perceived changes as the pilot progressed. Additionally, survey questions were used to test how widespread an issue or opinion identified in interview data might be among all participants. A random sample of pilot and control school parents received surveys in English and Spanish. All surveys were returned directly to an independent scanning service.

Individual and Group Interviews

The surveys were supplemented by a series of confidential individual and/or group interviews of pilot and control teachers and principals, board members, district staff members, Design Team members, Association leaders, parents and a range of external stakeholders and funders. The interview protocols were designed to gain perspective on the impact of the pilot and changes in the impact, as well as an understanding of how individuals were experiencing elements of the pilot such as objective setting. Additionally, in seeking to identify factors or conditions that were potentially contributing to or impeding the success of the pilot, teacher perceptions of the fairness and credibility of Pay for Performance were followed through the life of the pilot.

Student Achievement Data

The analyses of the student achievement data for the pilot and control schools were based on the district's two most commonly administered assessments, the *ITBS* and the *CSAP*. These data have been used to follow achievement over the four years of the pilot.

Documentary Data

Documentary or artifact data were used to gain greater perspective on areas such as school plans, teacher orientations, policy development, other initiatives, and internal and external communication related to the pilot. The most significant body of artifact data for this study were the teacher-developed objectives, which were read each of the four years and rated based on four quality criteria.

Observations

There were also observations of and participation in pilot implementation processes. CTAC was present each month at key planning meetings, both formal and informal, in order to gain an

understanding of the decision-making processes, the complications and methods of resolution during pilot implementation, as well as the ongoing development of the compensation plan. Although teaching was not a subject of this study per se, observations based on the *Performance-based Standards for Colorado Teachers* were conducted in the classrooms of sixteen pilot teachers selected as representative of the total participant group.

The analyses of these data constitute the substance and findings of the Pay for Performance study.

D. The Content of the Mid-Point and Final Reports

The negotiated agreements called for two reports of the results of the pilot. The mid-point report, *Pathway to Results: Pay for Performance in Denver,* was published in December 2001 and delineated the findings from the baseline year (1999-2000) and the subsequent year (2000-2001). This final report is based on data from all four years (1999-2003) of the pilot. It is possible to read the final report and understand the character and outcome of the pilot without having read the earlier report because the areas of inquiry remained constant as additional years of data and different types of data were added to the study. The aggregation of four years of data, the identification of trends, and the findings that emerged from the analyses of these data present a fuller picture with more longitudinal data than the earlier report.[4] However, the two reports are written as companions.

As the reader of this report will discover, the study of the Pay for Performance pilot is more than an examination and analysis of data. It is a story as well. Thus, within this report and alongside the evidence and findings, there is also a narrative. Like all stories it has beginnings, decisions, players, complications, resolutions, and results. A large body of the evidence supporting the findings can be found in the accumulated experience of participants who have told their stories each year to the researchers. One participant, a member of the Joint Task Force on Teacher Compensation, says this eloquently:

> "PFP is a story and it must be told right. We must get set up to accomplish the mission… Designing and implementing infrastructure is high art. We must recognize that it won't be perfect. We must create an environment and be allowed to recreate it over and over again. We are always going to be planning the perfect new system and good leadership can help make this transition from one iteration to the next. Each new attempt will have its strengths and weaknesses. But it's this process that allows for new staff growth and commitment."

The chapters that follow contain the analyses, the story, the process, and the results of the first iteration of Pay for Performance in Denver.

Chapter II describes in more detail the genesis of the pilot. The origins of the pilot were rooted in a unique collaboration between the Board of Education and the Denver Classroom Teachers Association. The design of the pilot is an outcome of the interests of the two parties rather than one based on an adopted model or an experimental research design. Additionally, this chapter considers the Denver pilot in the context of documented objections to merit or incentive pay plans of the last two decades.

Chapter III explains the research design of the study. Because this study uses a mixed-method design and because it was conducted in a large school district with an evolving educational program, the design is complex and multi-faceted. As the chapter shows, CTAC worked diligently with the school district and Design Team to ensure high quality research standards. The effort was not without its complications, which are explained in the chapter. Additionally, CTAC engaged the thinking of outside experts to address some of the statistical and research dilemmas that emerged. Deeper qualitative studies were added in the fourth year in order to broaden the understanding of and verify several findings.

Chapters IV and V explain the process used by teachers to set and measure objectives and the methodology used by CTAC to study these objectives. Since this was new ground not only for the teacher participants and the leaders of the

pilot but also for educational research, it required comprehensive methodologies. Because there is a significant relationship between the highest quality of the objectives written by teachers and the growth of (elementary) students on independent measures that has held up even as the number of high quality objectives written by teachers increased, objectives as an element of PFP inspire thought-provoking questions about teacher planning and practice.

Chapter VI discusses the four-year trends in the effects of the pilot on student achievement in the pilot and control schools. These are results from the *Iowa Test of Basic Skills* and the *Colorado Student Assessment Program*. The chapter discusses the utility of the two measures and the importance of being able to follow individual student growth over multiple years. Additionally, the reader will learn how CTAC worked to overcome some of the bias inherent in the pilot design and implementation.

Chapters VII and VIII look at the impact of the pilot on the Denver Public Schools, discussing the way in which the pilot has acted as a catalyst for change of the larger organization and also identifying the challenges to organizational alignment and systemic quality that implementing a pay for performance system entails. Most of the issues of fairness and credibility in PFP that were identified by teacher participants result from systemic weaknesses and gaps, the most glaring of which concern the adequacy of student assessments and professional development for teachers and principals.

Chapter IX contains recommendations for the Denver Public Schools and the Denver Classroom Teachers Association as they move to the next iteration of Pay for Performance.

Chapter X provides an analysis of the national implications of performance pay systems with a set of recommendations for districts and unions embarking on this type of reform and for foundations seeking to promote systemic change in American public education.

CHAPTER II

Pay for Performance

A. Introduction

By the end of the 1990s, an education accountability movement in the form of state standards and assessments, school report cards, and an emphasis on increased teacher effectiveness was in full swing nationally and in Colorado. While many of the reforms aimed at improving student achievement were showing uneven results, there were compelling data demonstrating that "differences in teacher effectiveness were a dominant factor affecting student gain."[1] Secondly, findings from other teacher effect studies identified teachers with strong verbal and math skills and deep content knowledge as having significantly positive effects on student achievement. As one study concludes, even if the findings are not sufficient to explain effective teaching fully, these two traits alone form a sufficient basis upon which to take action until further research advances a fuller explanation of teacher effectiveness.[2]

In Colorado, discussions at an August 1998 retreat of the Denver Public Schools Board of Education culminated in a paper wherein a cornerstone of the Board's vision was to "change the way teachers are paid." With the intention of linking teacher compensation to student achievement, the Board of Education established criteria for a plan that would be "fair, competitive and attractive to employees." Board members wanted, among other things, to eliminate automatic raises, link them to the achievement of specified goals, and create a compensation system that would lead to a focus on student achievement without teachers feeling competitive with one another. Board members also wanted to measure achievement in terms of individual student growth, or the value added by individual teachers. These interests became the basis of discussion and negotiation with the Denver Classroom Teachers Association. By the fall of 1999, the Denver Public Schools and the Association had committed contractually to a greater emphasis on results for students and teachers. They collaboratively sponsored a pilot designed to link teacher compensation and student achievement: Pay for Performance.

At the outset, the pilot was a momentous meeting of the minds between labor and management in a large urban school district on matters dear to the hearts of both parties. Both were interested in improved student achievement and professional performance. The Board of Education designated the pilot as one of the district's highest priorities; it was also one of the highest priorities of the Association. They both committed to the successful implementation of the pilot.

The agreement between the two parties also established new structures to advance the implementation of the pilot and develop a new compensation plan: (1) the Design Team, comprised of two Association appointees and two district appointees, which set to work immediately on designing and implementing the pilot; and later, (2) the Joint Task Force, comprised of appointees from the Association, district and community, which began to create the architecture for a new compensation system. Several other informal committees and partnerships from the district and the Association were essential to the success of the pilot. Ultimately, the interest and commitment of resources from private foundations helped make the full pilot and study possible.

Four years later, in the spring of 2003, as the pilot phase of Pay for Performance came to a close for the district and its teachers, the accountability environment in Denver, as in all districts in the country, had greatly intensified. The *Colorado Student Assessment Program (CSAP)* together with the No Child Left Behind Act (NCLB)—the 2001 reauthorization of the Elementary and Secondary Education Act that guides most federal education spending—have absorbed and escalated many of the components of the school accountability reforms of the previous decade, including standards-based education, highly qualified teachers for all students, research-based strategies, and a defined series of consequences for underperforming states, districts, and schools.

An outcome of Pay for Performance in Denver may prove to be, among other contributions, a serious consideration of how reform in teacher compensation systems can contribute to systemic accountability for student achievement.

B. Genesis of the Pilot in Denver

There are many perspectives on how the Pay for Performance pilot emerged from contract negotiations and what chances such a pilot had for success. Indeed, the definition of success varied among key players even as the pilot was underway. In the spring of 2000, one board member suggested that success meant "a compensation system based on outcomes and the performance of teachers." For another board member, success was seen as "whether we get it ultimately adopted, whether there is enough success to implement." A central administrator described success as "getting into the schools: What's the impact on kids? On the teacher's sense of purpose? On morale issues? On strategically and administratively dealing with PFP as an asset?" A teacher leader responded that success would be: "Curriculum and Instruction, and Assessment [two different departments] aligned. If they were aligned, that would be a success. A good salary schedule would be a success, but it is not what causes success. If student achievement goes up, then we are successful. If systems were put in place to support the sites, that would also be a success."

While the Board had designated teacher compensation and student performance as the cornerstone of its vision in 1998, the Association had a committee that had been studying pay for performance since 1994. During negotiations, Association leaders were not only aware of various systems of performance-based pay in Colorado and across the country, but were also concerned about the possibility that teachers might become the targets of an administrative fiat. One teacher leader noted at the time: "Oregon has imposed PFP through legislation. If Colorado imposes it, we'll be glad that we tried it out on our own and that we have data." Getting involved meant getting a voice, as another teacher leader reports: "[Teachers] wanted a voice . . . Now we have two members on the Design Team who are the leaders and who help teachers. We want to be a part of reform instead of just being the object of it."

With a program design that was the creation of negotiations and with so little lead time before

implementation got underway, the Pay for Performance pilot nonetheless proved resilient during the fall start-up. Partly this was because participation in the pilot was permissive, based on a faculty vote, and partly because teachers already had experience in setting annual objectives. They would get bonuses for "doing what they already do" and, as a fall back at the end of the pilot, they could vote for or against a permanent system that involved pay for performance. Further, "relationships among teachers [were] cohesive," according to a teacher leader. There was also a perception that, via the Design Team, teachers were leading the implementation of the pilot.

Altogether teachers maintained a significant piece of autonomy during the negotiations on the design of Pay for Performance. The leadership feel this was achieved by (1) basing performance-based compensation on teacher-set objectives; (2) limiting the scale of the implementation to 12 schools initially; (3) letting schools opt in; and (4) promoting the engagement of an independent outside organization to conduct the study.

The quick start-up did mean that many of the details of the design of PFP were worked out as the pilot was being implemented. It also meant that many of the central administrators and building principals, major stakeholders in a teacher pay for performance system, were not initially exposed to the concept of the pilot in a systematic way, leading to some of the implementation difficulties that emerged along the way.

A teacher leader described the implementation of the pilot: "The needs of PFP are out ahead of the district.... This has been like 'skunk work' since we have to make up everything as we go. We also have had trouble getting into someone else's backyard... [a central administrator] called PFP a "virus" that gets into every department in the district... all of which have their own agenda. The pilot is forcing speed in places where there have been no timelines, so we have had to build bridges and alliances [in order to get the needs of the pilot met]."

Since there were not a significant number of successful performance-based compensation models to follow, and since the two major teacher unions have been wary of performance pay, many of the sources of information and support that underpin the work of reform were not available to Denver participants. The pilot was breaking new ground.

C. Merit Compensation in Education

The idea of paying teachers according to their performance has a long history, yet actual attempts to do so have not successfully competed with the prevalent single salary schedule. This schedule is based on the accrual of teaching experience and the acquisition of education units. The attempt to pay teachers based on their performance or perceived performance has been traced back to 1710 in England where the practice became a part of the British Revised Educational Code in 1862. However, in the 1890s, it was removed based on the belief that it produced teaching to tests, rote learning, and cheating.[3] More recently, in the 1980s, a variety of incentive plans were implemented by state and local school agencies in order to improve the attractiveness and quality of careers in teaching, retain the best teachers and theoretically improve teacher performance and effectiveness. In 1986, a review of incentive programs intended to motivate teachers, identified three theories upon which such measures are based: (1) expectancy theory, that individuals will work for a valued reward, such as a bonus or promotion; (2) equity theory, that individuals will be dissatisfied based on feelings of unjust compensation for their efforts and accomplishments; and (3) job enrichment theory, that challenge and variety lead to greater productivity.[4] The first two theories support merit pay and career ladders, while the third one suggests that opportunities for new and different work should be structured into the workplace.

Even more recently, a study argues that it is time to break out of the single salary schedule and to change how teachers are paid. This position argues that merit pay systems have not lasted because (1) teachers are uncomfortable with the subjective judgments of administrators; (2) funding streams have not lasted; (3) merit pay conflicts with the collegiality that characterizes good schools; and (4) the education community has not had viable models upon which to construct new pay systems.[5]

The lack of models may soon be addressed as more school districts and states enter into pay for performance experiments and as prominent organizations promote initiatives that link teacher compensation to teacher effectiveness. "Investing in Teaching"—sponsored by the Business Roundtable, the National Association of Manufacturers, the U. S. Chamber of Commerce and the National Alliance of Business—has indicated support for a range of experiments with teacher compensation, including the Denver pilot.[6] Further, while many teacher unions at the local and national level have opposed direct linkages between compensation and student performance, some, including several districts represented in the Teacher Union Reform Network, have promoted union sponsorship of such initiatives.[7] Additionally, many local unions, as in Denver, have worked with boards and district managers to develop customized approaches to promote and reward teacher effectiveness.

There are differences in the approaches to performance pay. Some are based on teacher demonstration of skills and knowledge; others on the performance of groups of teachers or schools; and some, like Denver's Pay for Performance pilot, are based on the belief that student achievement is the bottom line and that part of compensation needs to connect directly to student results. Despite the increased numbers of experiments and the various approaches, there is yet to emerge a single approach that has demonstrated sustained success in supporting student learning in a large school district. Concomitantly, few comprehensive studies of the results of performance-based compensation experiments have been undertaken.

D. Potential of PFP in an Era of Accountability

In its final design, Pay for Performance focuses on teacher output—what students learn during their time with individual teachers. By emphasizing student growth, the design of the Denver pilot overcomes several of the objections raised about performance-based compensation. The design addresses these objections in the following manner:

- The teacher sets the objectives, either individually or with a group of colleagues, and with the approval of the building principal. At the end of the year, he or she provides evidence of attainment to that principal for validation. In this manner, the judgment of the administrator tends to be objective, based on student achievement data provided by the teacher.
- The teacher measures growth from where the student starts at the beginning of the year and then sets the growth target. Consequently, there is less likelihood that student factors (e.g., socioeconomic status, etc.) will affect the teacher's ability to earn a bonus or additional compensation.
- The teacher selects the measures, thereby decreasing the potential of teachers teaching to one monolithic test.
- The implementation of Pay for Performance has apparently contributed to collegiality rather than increasing competition. For example, the Spring 2001 survey findings indicated that 23.0% of pilot school teachers and principals believed that cooperation among teachers had increased; 66.7% indicated that it had stayed the same; and 10.3% indicated that it had declined. On the same survey, 9.4% of teachers and building principals indicated that competition among teachers had increased; 83.1% indicated that it had stayed the same; and 7.5% reported that it had declined. In Spring 2002, 39.8% reported that cooperation among teachers had improved; 51.3% that cooperation had stayed the same; and 8% that it had declined. In Spring 2003, 53.4% of respondents reported that PFP had had a positive impact on cooperation among teachers; 44.5% no impact; and 2.1% a negative impact.

Figure 2-1 compares several of the longstanding concerns about performance-based compensation to the Denver pilot.

The realities of pilot implementation of PFP have brought up legitimate issues of credibility and fairness among teachers, administrators, pilot leaders, and researchers that are discussed in this report. There remains a belief among pilot leaders and many pilot teachers that continued

FIG. 2-1

Areas of Concern About Performance-Based Pay

Objections/Issues/Concerns	Pay for Performance in Denver
1. Subjectivity of the administrator	Teacher develops and brings evidence of attainment to the administrator who has already collaborated in the development of the objectives
2. Differences in student background	Student growth is measured in annual terms
3. Teaching to the test	Teachers select their own measures so, if they are teaching to the test, it is not one monolithic test
4. Increased teacher competitiveness	53% report a positive impact on cooperation; 44% no impact; and 2% a negative impact
5. Unstable funding streams	The Joint Task Force on Teacher Compensation began to address this issue
6. Lack of reliable and valid measures of student achievement for teacher use	Measures used by PFP too open-ended to be tested in the study

refinements in the design of and support for Pay for Performance can address such issues. As discussed in Chapter VII, teachers have regularly been making recommendations for improvements in the design. A pilot teacher indicates, "We need uniform procedures for objectives. The timeline didn't make sense. We started the school year in mid-August, goals written by mid-October, reviewed by November or December. It should have been done earlier in the year. I didn't get the 'OK' on my objectives until December." Another pilot teacher says, "The one thing that bothers me about PFP is that the objectives [for] non-academic teachers (music, PE, library, etc.) do not compare to what regular classroom teachers are doing. It's not equitable." Another pilot teacher states, "We have to find a way to distinguish between school politics and the pecking order in schools and actual teacher performance. Right now those are meshed together. Right now the leadership has too much decision-making power and that causes anxiety."

As Pay for Performance moves to another iteration in Denver, issues of credibility and fairness—identified during the implementation of the pilot—will continue to provide a basis for needed improvements. However, the fundamental design—examining progress based on the attainment of teacher-set objectives—will likely have staying power.

E. The Bargaining Agreement

Initial Terms

Appendix E of the Bargaining Agreement between the Board and the Association contains the terms of Pay for Performance. Key features of the agreement include:

- Setting a two-year term for the pilot.
- Commissioning the Design Team, comprised of two teachers selected by the Association President and two administrators selected by the Superintendent, and releasing all four members from their present duties.
- Charging the Design Team with designing, overseeing, implementing, and evaluating the pilot, including the authorization to seek an outside evaluator.
- Defining the terms for participation in the pilot (initially 12 elementary and three middle schools), where 85% of the faculty had voted to participate.
- Establishing the financial terms which included: (1) maintenance of the salary schedule; (2) in year one a $500 stipend for participation and $500 bonus for each objective obtained; and (3) in year two a $750 bonus for each objective met.
- Setting up three approaches that teachers in participating schools would use to measure progress: (1) the *Iowa Test of Basic Skills,* a nationally normed test; (2) teacher-developed criterion-referenced tests or other teacher-

developed measures; or (3) increases in teacher knowledge and skill.

- Establishing dates for the Design Team to report to the Board and the Association.

Revised Terms

The agreement between the two parties allowed for mid-course corrections and revisions to the pilot. Most of the substantive modifications to the design of the pilot occurred during the first year and were the result of efforts to make a greater level of science possible in the study. Based on concerns presented by the Design Team and CTAC, and with the collaborative support of the Board of Education, the Association and external parties, several revisions to Appendix E at the policy and operational levels were worked out in January 2000. These are detailed in the mid-point report, *Pathway to Results.* They included:

- Extending the pilot to a period of four years in order to establish a baseline year and to provide more than two years of student achievement data for a more accurate identification of student achievement trends.
- Establishing new reporting dates and products, including a mid-point report and a final report to be provided by CTAC.
- Defining the baseline year for study purposes as 1999-2000.
- Changing the threshold for faculty votes to participate from 85% to 67%.
- Establishing the need for a group of control schools.

In June of 2000, two additional challenges were addressed through another collaboration of the key parties: (1) the need for a written statement of purpose to provide direction and clarity to the pilot; and (2) the need for a vehicle to directly address the development of a new compensation system.

The formal Statement of Purpose follows:

> *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 teachers' 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 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 later established with the charge to "develop and analyze a model salary system for appropriate teacher pay for performance in the Denver Public Schools." This group is comprised of representatives of the Association, administration, and the community at large.

Although numerous corrections occurred after the presentation of the mid-point report, which is discussed later, most revisions to the basic structure of the pilot occurred within the first two years and with the consent of all parties. Undertaking revisions such as the ones outlined above showed the commitment of the sponsors to the concept of pay for performance. They were prepared to make corrections to strengthen the implementation and the study of the pilot.

F. The Role of Community Training and Assistance Center

In November of 1999, CTAC was selected to fill a dual role with regard to Pay for Performance. First, it provided technical assistance to assure pilot quality and integrity. Second, it conducted the study of the impact of the pilot.

In meeting Denver's expectations, CTAC became a participant-observer developing a case study of the implementation of Pay for Performance. Specific recommendations or technical assistance have the potential of introducing bias into a study. While this potential cannot be eliminated, CTAC has taken the steps identified as appropriate for this form of research.[8] First, the study draws on multiple sources and has clearly identified those sources in describing what has taken place and in drawing conclusions. Second, the two reports have described this relationship to pilot participants and readers. Third, the study identifies the sources of and routes to conclusions. The quantitative data (student achievement data and survey results) are a matter of record; the qualitative data (interviews, focus groups, observations, and artifact analyses) have been collected and analyzed in written format and, for reasons of participant confidentiality, are less accessible.

The involvement of CTAC staff has also included: (1) observation of district activities and regular pilot leadership meetings; (2) assistance to Denver assessment, planning and data professionals in the maintenance of student and teacher databases; and (3) reports of annual collections of data (without interpretation) to the Design Team, Board of Education, the Association, superintendent, funders, and other interested stakeholders for their use in determining the status of the pilot.

The identification of issues and subsequent recommendations are matters of interpretation and should be seen in light of CTAC's complete role. However, these findings and recommendations have been formally submitted to the superintendent, pilot leaders and sponsors in management letters, as well as presented in the mid-point and final reports. The superintendent, pilot leadership and sponsors determined the level and quality of response to the findings and recommendations. CTAC played no role in the initial design or structure of the pilot, nor in any of the related negotiations, but it has been a close observer of subsequent activity.

Denver leadership wanted to know not just what happened, but also *why* and what needed to be considered next. Responding required a certain level and form of involvement. Context is critically important both in interpreting outcomes and in determining next steps. Therefore, CTAC's active involvement in the pilot implementation was both a requirement and a strength of this study.

G. Summary

As this chapter shows, the Pay for Performance pilot in Denver has benefited from the spirit of collaboration and innovation demonstrated by the Denver Public Schools Board of Education and the Denver Classroom Teachers Association. By establishing a pilot where the potential growth of student achievement and teacher professionalism could be explored and by removing the concept of pay for performance from the level of ideological discourse to a more scientific discourse, the district has created a pilot and study of merit. The impact of pay for performance and its potential for increasing student achievement and teacher professionalism will be better understood not only by Denver, but by others who are interested in this type of reform.

Of all of the current experiments in teacher accountability, compensating teachers based, in part, on student performance is among the most controversial. Based on failed attempts of the past, and legitimate concerns of teachers about the basis of their compensation, the concept enjoys both strong support and strong resistance in the education community. Pay for Performance in Denver, as shown in this chapter, addresses several of the criticisms of this type of approach. It also reveals issues and challenges that still need to be addressed. As a method of reform, support and accountability, Pay for Performance cannot be described as an effort to "tinker around the edges." It goes to the heart of a school district's systems in order to make serious reforms and accountability a way of life in a large district.

CHAPTER III

Research Design

A. Introduction

The primary purpose of this study is to examine the impact of the Pay for Performance pilot. As noted in the district's formal Statement of Purpose: "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."[1] A central question is whether growth on measures of student achievement can be linked to teacher performance in the Pay for Performance pilot. For this reason, the study focuses on changes in student achievement within the pilot schools and between pilot and control schools, as well as student growth associated with teacher objectives. Additionally, the study examines the nature and influence of teacher objective setting in detail; the impact of school, teacher, and student factors; and the differential impact of the pilot implementation on school and district-wide practices as perceived by teachers, administrators, and parents. Finally, the study examines the institutional factors that have affected implementation. The study is not an evaluation of the pilot. Rather, it is a much broader examination of the results of the pilot using both quantitative and qualitative measures, addressing questions of immediate impact and issues of policy making appropriate to Denver.

This chapter discusses the research design of the study, including research questions, methods of data collection, and methodological choices and rationales, along with dilemmas that arose in the use of district data.

B. Research Design

In order to answer the questions raised by the pilot, a mixed-method design was used, combining the results of quantitative data (e.g., student achievement data from the *Iowa Test of Basic Skills (ITBS)* and the *Colorado Student Assessment Program (CSAP)* and perceptual data from survey questionnaires) and qualitative data (i.e., interviews, focus groups, artifacts, and observations) collected in all four years of the pilot. The use of more than one method to study the same

phenomenon, called triangulation, strengthens the validity of the results.

Another advantage of the mixed-method approach is that it may lead researchers to modify or expand the design and/or data collection methods. This action may occur when inconsistencies and discrepancies have been uncovered. For example, some interview and survey data indicated that teachers were not changing their teaching practices, but still other findings, such as the increased quality of objectives and the increased focus on student achievement, suggested that teachers were doing something differently. To further investigate anecdotal evidence of changes in behavior, a deeper qualitative study, including focus groups and classroom observations, was implemented. This study and other such responses to unexpected events are described in detail in the chapters that follow.

Research Questions

Building on the pilot infrastructure established by the Denver Public Schools, the Denver Classroom Teachers Association and the Design Team, the research study was designed to address four overarching research questions:

1. What is the impact of Pay for Performance on student achievement?

Providing additional compensation based on student performance is what most distinguishes Pay for Performance in Denver from efforts attempted in other school districts in the United States and abroad. The study examines: (a) the changes in student achievement which have occurred at the pilot schools, and (b) how these changes in achievement at the pilot schools compare to those in control schools. The analyses of student achievement data also incorporate student, school, or teacher factors.

2. What is the impact of the teacher-developed objectives?

At the pilot school level, the objectives set by individual teachers were the centerpiece of Pay for Performance. Teachers received additional compensation only when they met their objectives. The study examines the relationship between teachers' meeting their objectives based on their own assessments and actual increases in student achievement based on independent measures. Further, the study explores the substance of the objectives—as based on a four-trait rubric—in order to determine whether the quality of objectives can be associated with increases in student achievement.

Other questions are also addressed. For example, how have objectives changed since the pilot began? How do student achievement results compare to teacher objective ratings? How has the objective setting process impacted teacher practice? How has the objective setting process worked for special subject teachers (e.g., art, music, physical education), special educators and specialists (e.g., nurses, psychologists, speech and language specialists)? What issues arose for various pilot participants? What kinds of supports were received?

3. To what extent do school, teacher and student factors impact and/or explain student achievement?

The implementation of Pay for Performance occurred within a functioning school system where other site level factors, beyond teacher compensation, may have influenced student achievement. The study attempted to identify to the greatest degree possible those site level factors that contribute to, and may prove to enhance or impede, the achievement of students or the effectiveness of the pilot.

There are significant differences among the student populations at pilot and control schools as well as teacher factors, such as years of experience. Are there differences in student or teacher demographics that explain student achievement? Are there differences in the achievement history of pilot and control schools? How do we control for these variables in assessing the impact of the pilot on student achievement?

4. What broader institutional factors have influenced the implementation of the pilot, and how have these factors affected pilot results?

The pilot exists within a broader institutional context, a large, urban school district, that must respond to challenges from both within and without and over a four-year period. Pay for Performance, as an initiative with significant systemic implications, was limited by the ability of the

district to implement a major new initiative. The study analyzes the institutional factors that have had the most marked impact on the pilot. The study examines policy and operational decisions, support structures and assignments, mid-course corrections and related interventions, the perceptions of different constituencies, and the lessons that have emerged during the implementation of the pilot.

What institutional factors influenced the implementation and outcomes of the pilot? What systemic barriers confronted the implementation of the pilot? What factors outside of the district, such as state and national initiatives, affected the pilot?

Selection of Pilot Schools

In the fall of 1999, the Design Team held sector meetings and more than a dozen school visits to promote participation in the Pay for Performance pilot. Elementary and middle schools voted to determine if their schools would participate. The original DPS/DCTA Agreement required 85% of the faculty to vote in favor of participation in order for a school to be included in the pilot. The twelve elementary schools that met this threshold comprised the original pilot schools. The threshold was later lowered to 67% and additional elections were held.

In June 2000 or beginning in the 2000-2001 school year, Horace Mann Middle School became the first secondary school to join the pilot. The original 12 schools were given the opportunity to withdraw from the pilot in December 2000. At that time, Smith Renaissance Elementary School chose to withdraw. In the third year of the pilot, another elementary school (Philips), another middle school (Lake) and two high schools (Manual and Thomas Jefferson) joined the pilot. *Figure 3-1* shows the participation of schools by year and explains the fact that analyses and discussions may refer to different numbers of schools in different years of the study.

In the 2002-2003 school year, Manual High School officially split into three smaller schools: Arts & Culture, Millennium Quest, and Leadership Academy. For analytical purposes, the three schools were treated as a single entity throughout the study. An analysis of the new school populations shows that students selected or were selected into the new small schools such that ability groups are concentrated rather than diffused throughout each of the three smaller schools. Nonetheless, it is not possible to assess the impact of the pilot accurately at Manual High School independent of the change in school structure. For this reason, achievement results are presented separately for Manual and Thomas Jefferson High Schools.

Selection of Control Schools and Related Issues

The study design included comparison schools to control for (1) the effects of contemporary history, and the effects of selection-maturation interaction. In the first case, the inclusion of control groups limits the possibility that contemporaneous events account for the change observed achievement in the pilot schools since both groups have experienced the particular event (e.g., the tragedy of September 11). In the latter case, the use of control groups limits the likelihood that an unmeasured factor not reflected in the pre-test, but operating to contaminate the post-test data (e.g., changes in the administration and importance of *CSAP* over the life of the pilot).

In the original pilot proposal in January 2000, CTAC requested control elementary schools to be used as a non-treatment comparison group. Three schools were to be selected for each pilot elementary school. In January 2001, the district identified the elementary control schools in the following manner: the schools were chosen to "match" each pilot school based on three criteria: (1) the percent of free/reduced lunch students; (2) the percent of English language learners; and (3) school size/enrollment (where possible). District assessment staff determined that the first two criteria were the most important and the third was matched where possible. In the case of one school, Smith, although the district included Smith as a control, CTAC included only the baseline and the one year in which Smith participated in the pilot and did not use Smith as a control school because there may have been lingering effects from the pilot.) The district determined that all middle and high schools were to serve as control schools at the secondary level. The schools designated by the district are listed in *Figure 3-2*. The Career Education Center was not used in the study as a control

FIG. 3-1

Participation of Pilot Schools by Years in Pay for Performance Pilot

School	1999-2000	2000-2001	2001-2002	2002-2003
Centennial	√	√	√	√
Colfax	√	√	√	√
Columbian	√	√	√	√
Cory	√	√	√	√
Edison	√	√	√	√
Ellis	√	√	√	√
Fairview	√	√	√	√
Mitchell	√	√	√	√
Oakland	√	√	√	√
Philips			√	√
Smith Renaissance	√	√		
Southmoor	√	√	√	√
Traylor Fundamental	√	√	√	√
Lake Middle School			√	√
Horace Mann Middle School		√	√	√
Manual High School			√	√
Thomas Jefferson High School			√	√

school because it is a non-traditional school and because its testing rates were low.

Unfortunately, though the selection of control schools for the pilot appears demographically reasonable, previous achievement in control schools was not a factor in their selection and comparability to the pilots. The controls had lower test scores than the pilots on the Spring 1999 administration of the *ITBS*. This fact makes it more difficult to detect a PFP positive result because higher performing schools will tend to regress downward toward the mean and lower performing schools will often tend to rise toward the mean over time. Secondly, there is no way to disentangle the effect of the pilot from the characteristics that are associated with teachers who self-selected into the treatment group. Differences between pilot and controls could be due to whatever factors caused teachers to vote to be included or not in the pilot.

A second complication arose when, early in 2001, schools were advised by the district that administration of the *ITBS* was optional. However, this advisement was later retracted for pilot and control schools since the *ITBS* was one of the two standardized measures being used in the study to assess student achievement and the district's only norm-referenced longitudinal measure.

A few weeks before the test was to be administered, control schools were informed, some for the first time, that they were designated as control schools and, as such, would have to continue to give the *ITBS* each spring for the duration of the pilot. Some schools had not planned to administer the test, and so issues arose later regarding low testing rates. Testing rates will be discussed further in Chapter VI.

Mid-point Changes to the Design of the Pilot

At the end of the first year, a review of the original design was conducted to gain direction for the Pay for Performance pilot as it was developing and to answer such questions as: Which of the activities or strategies are aiding the participants to move toward the goals of the pilot? What barriers have been encountered and what needs to occur in order to overcome these barriers? Some changes were identified as early as June 2000:

- Extending the pilot to a period of four years.
- Defining the baseline year for study purposes.
- Changing the threshold for faculty votes to participate from the initial of 85% to 67%.
- Establishing the need for a group of control schools for study purposes.

In December 2001, the mid-point report was presented to the Denver educational community. It defined the impact of the pilot at the halfway mark and delineated changes needed to increase the effectiveness of the pilot and barriers yet to be addressed. Changes in the pilot design and implementation that resulted from the recommendations made at the time of the earlier report include the following (which are described in more detail in the coming chapters of this report):

- Developing learning content explicitly in the objectives.
- Addressing the fairness related to special subject teachers, special education teachers, and specialists.
- Providing teachers with more support in objective setting.
- Integrating the three approaches that were originally part of the pilot design.

FIG. 3-2

Elementary, Middle and High School Controls

Elementary School Controls		
Amesse	Goldrick	Remington
Asbury	Greenlee	Rosedale
Ashley	Gust	Samuels
Bromwell	Holm	Schmitt
Cheltenham	Kaiser	Slavens
Doull	Lincoln	Steck
Ebert	Maxwell	Steele
Fallis	McGlone	Teller
Force	McMeen	University Park
Garden Place	Montclair	Valverde
Gilpin	Moore	Whittier
Godsman	Newlon	
Middle School Controls		
Baker	Hamilton	Merrill
Career Education Center	Henry	Morey
Cole	Hill	Place
Denver Schools of the Arts	Kepner	Rishel
Gove	Kunsmiller	Skinner
Grant	Martin Luther King	Smiley
High School Controls		
Abraham Lincoln	George Washington	North
Career Education Center	John F. Kennedy	South
Denver School of the Arts	Montbello	West
East		

Another critical change occurred in June 2002 when the supervision and reporting of the Design Team and the pilot was transferred to the district's chief academic officer. While this change gave the Design Team and the pilot a more mainstream relationship within the district and moved it into the center of the instructional program, it also created some confusion between the pilot and newer instituted initiatives such as the district's new literacy initiative.

Dilemmas and Caveats

Interpretation of the quantitative results is limited by a number of the study's design and implementation features. Though some of these were amenable to mid-course corrections, others were not. Entry into the pilot was by self-selection, by vote of the teachers in each school. This method of selecting pilot schools ensured greater teacher cooperation, but also limited the applicability of the pilot findings to other settings. In a setting where teachers are not given the choice to participate, the outcomes could be quite different. Self-selection also leaves the possibility that an unmeasured or 'latent' characteristic of the pilot schools both led the schools to select into the study and caused any differences in student achievement noted between pilots and controls.

The use of the Online Assessment Score Information System (OASIS) and the Web-Based Objective Setting software by non-pilot schools for the purpose of writing objectives. Effectively, several control schools used PFP protocols and processes or modified forms of them, complicating the pilot-control relationship for the purposes of the study.

Testing rates (the number of students assessed annually) were not well monitored within the district, leading to lower than desirable numbers of students tested in some schools and years. For these reasons, the student achievement results of the pilot must be interpreted with caution.

C. Impact of Pay for Performance on 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 quality of objectives, have had on achievement.

In September 2000 the Design Team, in conjunction with the Assessment and Testing Department created an Assessment Matrix which identified 13 district-approved assessments for use in the different elementary grades, including the *Iowa Test of Basic Skills (ITBS),* parts of the *Colorado Student Assessment Program (CSAP),* and the *6+1 Trait Writing Sample (Six-Trait).* Measures for younger children were encompassed within the *Colorado Basic Literacy Act (CBLA)* and Title One/Grade Level Math[2]. Because all pilot school teachers are involved, including classroom teachers, special subject teachers (e.g., physical education, gifted/talented, music, art), special education teachers, and support services providers (e.g., psychologists, nurses, social workers, speech and language specialists), many different measures have actually been utilized in teachers' objectives. In a June 2000 report, the Design Team indicated that 116 different assessments were used by at least one teacher.

With the integration of the approaches and the inclusion of middle and high schools in the pilot, the number of different assessments grew substantially with a great many teachers creating their own tests. The assessments listed by teachers in measuring their objectives fall, for the most part, into three general categories: (1) assessments named in the district assessment matrix and unit tests which accompany text books; (2) assessments in a much looser sense such as attendance log, vocabulary list, formal lab reports, research paper or body chart word list; and (3) teacher-made or unspecified measures (e.g., pre- and post-tests, teacher's rubric, informal tally, oral and written tests).

In 2002-2003, a total of 1,260 objectives were reviewed by CTAC. Of these, 166 different "assessments" fell into the first two categories. A total of 471 assessments listed fell into the third category or 38% of the total objectives. A further breakdown indicated that of the 630 teachers writing objectives, 256 used some form of "teacher-made test" at least once (41%), while 146 teachers listed "teacher-made test" in both objectives (23%).

The level of effort necessary to analyze the entire set of assessments is beyond the scope of the present study. It is the task of the teacher and the principal in determining if the teacher did or did not meet their objectives. For the purposes of this study, three assessments were originally designated for analysis; namely, the *ITBS,* the *CSAP* and *Six-Trait.* At the beginning of the third year of the pilot, the district dropped *Six-Trait* from the district lexicon. It has been deleted from

the assessments analyzed in the study, although it was discussed in the mid-point report and many teachers (10.5% of the 630 teachers in 2002-2003) continue to use it in their objectives.

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. Initially the state's assessment, *CSAP,* which was designed for other purposes and which did not provide a mechanism for pre- and post-testing of an individual child, was less appropriate for objective setting and did not lend itself to the type of analyses one would prefer to use in a comprehensive study such as this one.

At the time of the mid-point report, the Colorado Department of Education indicated that in the future, it would be possible to examine reading scores from year to year through vertical scaling. However, when that report was prepared, *CSAP* could only be used for grade level comparisons and not to assess change at the individual student level. Significant changes have occurred in this assessment and will be described later in this chapter.

Description of Assessments

Iowa Test of Basic Skills (ITBS)

The *ITBS,* developed by the Riverside Publishing Company (1993), is a norm-referenced achievement battery composed of tests in several subject areas. The district administers and scores these tests. 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, scale scores and norms. In Denver, different grades were required to take different subtests from year to year, preventing comparisons of some grades and tests from one year to the next.

In the 2000-2001 school year, DPS decided to use the *ITBS* as the overall measure to compare academic achievement in the Pay for Performance pilot. First, it can be used to measure student growth; that is, a student's score in third grade can be compared to their score in fourth grade and that comparison can be used to draw inferences about how much he/she learned.
Second, the district had extensive longitudinal data from these tests that allowed for trends to be examined from before the pilot began. Third, the tests are a more comprehensive battery. At the beginning of the pilot, the district's testing program required that "all students in grades 1, 3, 4, 6, and 7 must take, at a minimum, the Reading section . . ." and "students in grades 2, 5, 8, and 11 must take all the subtests" each spring.[3]

As discussed, schools were advised in Spring 2001 that the spring administration of the *ITBS* was optional. After discussions with the Board, central administration and the Design Team, it was clear that in order to complete the research study, it was imperative that pilot and control schools continue to administer these tests until the end of the pilot. These designated schools were advised by the district that they would need to continue administering these tests until Spring 2003. This was met with consternation by some principals, and while some schools did continue to administer to almost all students, others appear to have administered on a more selective basis. This fact may have created some unintended effects discussed further in Chapter VI.

It should also be noted that some students are excluded from taking the *ITBS* at the principal's discretion. This discretion is not based on a set of rules and may be exercised differently at each school. Also, in setting their objectives, teachers may exclude students who do not meet certain criteria from their growth targets; for example, they may have entered a teacher's classroom midyear, or have been chronically absent. Since these factors do not appear in the district database, they could not be considered in this analysis.

Colorado Student Assessment Program (CSAP)

CSAP was developed for the State of Colorado by CTB/McGraw-Hill and was first administered in 1997. These tests are based on the Colorado Model Content Standards and were originally intended 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 items selection process, and to place both multiple-choice items and constructed response items on the same scale.

When *CSAP* performance levels were established from 1997 to 2000, the Bookmarking Standard Setting process was used for every grade level and content area. Scale score cut-points were set that defined four performance levels—Unsatisfactory, Partially Proficient, Proficient, and Advanced.

Use of the *CSAP* was problematic in the early years of the pilot for three reasons. First, the tests were not useful for measuring student growth because they were not given in contiguous years; secondly, they have been phased-in in a staggered fashion (see *Figure 3-3*) with one or two tests introduced per year since 1997. Finally, the battery of tests was not comprehensive because it did not offer grade-by-grade data in two content areas from grades 2-11.

The *CSAP* environment changed significantly after the release of Senate Bill 00-186 which requires that the *CSAP* Reading tests be administered in contiguous grades and reported on one common, vertical score scale. An additional influence on the nature and purpose of the *CSAP* was the introduction of the No Child Left Behind Act of 2001 (NCLB), which required that all students be assessed longitudinally in reading and math in grades 3-8 so that their progress can be measured against state standards. Under the NCLB provisions, annual tests in reading and math must be in place by the 2005-2006 schools year; however, 2002 is the base year for determining adequate yearly progress and efforts to make adequate yearly progress began immediately in Colorado.

The Colorado testing program was significantly changed as a result of these two events as well as from requests from districts within the state. In a letter dated April 24, 2001, districts were advised by the Colorado Department of Education, Student Assessment Unit of the new scaling procedures and the expanded testing schedule to be implemented by the state department. These new procedures included a vertical rescaling of all tests administered since 1997 across all grades and content areas. The schedule of *CSAP* administrations during the life of the pilot is shown in *Figure 3-3*.

Quantitative Analysis Methodologies

School achievement data is hierarchical in nature. Students are grouped by classroom, grade, and school. At each level of the hierarchy, student scores are correlated. In addition, each student's scores are correlated over time.

Two-stage hierarchical linear modeling (HLM) makes it possible to account for the correlation within the school organizational structure. Because classroom level data was not available for the baseline year, the student achievement analysis employs a two-stage model, grouping students within schools. The two-stage HLM models allow each school to have a different intercept at baseline.

Individual growth modeling (IGM) extends the two-stage HLM model to take into account the correlation in student scores over time. IGM also uses a two-stage design to account for correlation within schools. In addition, the IGM model allows each student to have an intercept and slope: the intercept represents baseline achievement level and the slope represents the student's rate of growth over time. Details on the specification of the achievement models are found in Chapter VI.

D. Quality of Objectives

In the pilot schools, each teacher wrote two objectives. These were approved by the principal and formed the basis for evaluating classroom results. Objective setting is seen as a central component, if not the foundation, of the pilot.

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 database, the district scope and sequence (which contains subject standards for grades K-12), and the elements listed on the form provided by the Design Team to teachers. Four levels of performance were established as a way to rate

FIG. 3-3

Schedule of *CSAP* Administrations by Content Area, Grade and Year

		Grade							
Content Area	**Year**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Reading	99	√	√			√			
	00	√	√			√			
	01	√	√	√	√	√	√	√	√
	02	√	√	√	√	√	√	√	√
	03	√	√	√	√	√	√	√	√
Math	99			√					
	00						√		
	01			√			√		√
	02			√			√		√
	03			√			√		√
Writing	99		√			√			
	00		√			√			
	01		√			√			√
	02	√	√	√	√	√	√	√	√
	03	√	√	√	√	√	√	√	√

individual objectives. The levels of performance are as follows:

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

All objectives were read holistically and scored by multiple readers. *Figure 3-4* provides a breakdown of the number of objectives read over the four years of the pilot. In the first year of the pilot objectives were not yet in an electronic format and many of the objectives that were sent to CTAC for analysis and review were incomplete or duplicates. This resulted in the large number of unrated objectives. In later years more complete rating of objectives was possible due to the introduction of the Web-Based Objective Setting software created by the district. There were still a limited number of objectives that were duplicates or incomplete. A complete discussion and analysis of the rubric and the objectives are described in detail in Chapters IV and V.

Ultimately, the study used several sets of data to evaluate overall objectives quality: (1) rubric levels for each teacher's objectives over four years, 1999-2003; (2) the summary of met/not met objectives over four years, 1999-2003; (3) a comparison of objectives to the school plans in 2000-2001 and 2002-2003; (4) a comparison of pilot school objectives to control school goals, 2000-2001 and 2002-2003; and (5) achievement data on the *ITBS* and *CSAP* administered to all pilot schools for 1999-2003.

Objectives Met or Not Met

Over the course of the pilot, more than 4,000 objectives have been read and reviewed by multiple experts at CTAC. This review generally takes place in March and includes all objectives delivered by the Design Team. Two situations have caused the numbers reported by the Design Team and the numbers reported in the study to vary:

(1) teachers' objectives were not submitted and approved before March but were included in the district's report of met/not met objectives because they were approved before the end of the school year; and (2) teachers who have had their objectives read and reviewed left the district or moved to a non-pilot school and were not included in the end-of-year payout. *Figure 3-5* presents the numbers of objectives met and not met over the four years of the pilot as reported by the Design Team.

E. School, Teacher, and Student Factors

School, teacher, and student characteristics were collected for use in the quantitative analyses. They are used in the models to control for differences in school populations and characteristics between pilot and control schools.

School characteristics were collected from the school report cards. The factors selected for the analysis include number of years the principal has been at the school, percent of students who are English language learners, percent of students receiving free or reduced-price lunch, percent of students with a disability, percent of teachers who are not fully licensed, and total enrollment. All of these factors were centered at the mean at the elementary level, the middle school level, and the high school level. This makes it possible to interpret the coefficients in the achievement models relative to an average school.

Teacher characteristics were collected from the district human resource files. Chosen for the analysis were degree (bachelor's, master's, or doctorate degree) and years of experience in the Denver schools. The study determined which teachers were part of the Teacher-in-Residence program and included this information as well.

FIG. 3-4

Objectives Read and Rated by School

School	1999-2000	2000-2001	2001-2002	2002-2003	Total
Centennial	70	76	76	76	298
Colfax	50	52	54	52	208
Columbian	46	32	38	44	160
Cory	54	50	50	56	210
Edison	58	64	64	60	246
Ellis	70	68	70	72	280
Fairview	54	62	56	62	234
Mitchell	66	60	72	74	272
Oakland	70	70	78	82	300
Philips			58	54	112
Smith	70	66			136
Southmoor	20	34	40	44	138
Traylor	56	60	64	62	242
Horace Mann Middle School		94	108	92	294
Lake Middle School			132	120	252
Manual High School			168	168	336
Thomas Jefferson High School			152	142	294
Total	684	788	1,280	1,260	4,012

Student characteristics were obtained from the student demographic files kept by the district. Included in the analysis are grade, race/ethnicity (Native American, Black, Asian, Hispanic or White), any disability, English proficiency, grade retention, gender, and socioeconomic status. The study categorized students as non-proficient in English, bilingual, or English-speaker-only based on acombination of home language and socioeconomic status codes which describe a student's progress in learning English. SES is categorized as low (e.g., ever received free or reduced lunch) or high (e.g., never received free or reduced lunch). Student characteristics were examined over time and missing data were filled in based on the student's characteristics in contiguous years.

F. Impact on Teachers and Other Stakeholders

Purpose and Types of Qualitative Data

As part of the overall mixed-method design of the study, qualitative and quantitative methodologies were used to ascertain the impact of the Pay for Performance pilot on pilot teachers and other stakeholders. Surveys were sent to teachers, school administrators and parents. Individual interviews with board members, association leaders, central administrators, external community members, parents and a random sample of teachers and principals were conducted each spring.

In the first two years of the pilot, surveys and interviews were used to determine the level of awareness of the pilot, its goals and expectations as viewed by teachers and others in the district and the community. In the last two years of the pilot, these methods were used to explore perceptions of the impact of the pilot on various aspects of the district, including student achievement, professional development, the objective setting process and perceptions of a new compensation system based in part on student achievement.

Surveys

Over the course of the pilot, CTAC conducted surveys each spring of pilot school teachers and staff (2000-2003), control school teachers and staff (2001-2003) and pilot and control school parents (2001-2003). All pilot school teachers and staff, who participated in the pilot by submitting objectives, as well as the principal, received confidential surveys.

In the first year, the Design Team followed up with the schools to assure a strong response since this was to be a baseline for the study. In the case of the control schools, teachers and staff were sampled randomly from files provided by the district's Human Resource Department based on the size of the school (i.e., eight for small schools; 13 for mid-sized schools; and 27 for large schools). Surveys were also sent to the principals at each of the control schools. Respondents were

FIG. 3-5

Number and Percent of Objectives Met

Participants	1999-2000	2000-2001	2001-2002	2002-2003
12 Elementary Schools 342 Teachers/684 Objectives	629 Met 92.0%			
12 Elementary Schools 1 Middle School 421 Teachers/842 Objectives		770 Met 91.4%		
12 Elementary Schools 2 Middle Schools; 2 High Schools 635 Teachers/1270 Objectives			1113 Met 87.6%	
12 Elementary Schools 2 Middle Schools; 2 High Schools 644 Teachers/1288 Objectives				1288 Met 91.3%

directed to mail their completed surveys directly to the scanning center in postage paid, pre-addressed envelopes.

Random samples of parents from both pilot and control schools were sent surveys in the last three years of the pilot. Because CTAC did not have access to student names, the parent samples were drawn randomly using transformed student identification numbers which were then sent to the district which mailed the questionnaires addressed "To the Parents of…". English and Spanish versions were sent. All surveys were confidential. Each year between 300 and 400 of the surveys were returned by the post office as undeliverable. *Figure 3-6* presents a breakdown of the number of surveys sent and the number of usable surveys received for the four years of the pilot.

In the first two years of the pilot, the focus was mainly on the goals and expectations of the pilot as well as project support and project impact. This was the case with both the first year of the pilot when only pilot teachers and administrators were surveyed, and the second year when pilot and control teachers and administrators were surveyed. Parents, in the second year of the pilot, were asked similar questions. Beginning with the third year of the pilot, survey questions dealt with changes over the years of the pilot, and perceived impact of the pilot on changes in classrooms, schools, the district, and the compensation of teachers. Parents were also asked to respond to questions regarding the compensation of teachers and its relationship to student achievement.

Figure 3-7 provides a breakdown of the respondent groups across the four years of the surveys.

Individual and Group Interviews

Over the four years of the pilot, more than 600 individual interviews were conducted with pilot participants and other stakeholders. The range of interview subjects included members of the Board of Education, Denver Classroom Teachers Association leaders, central administration, external community members and funders, Design Team members, other site staff, principals, teachers and parents. *Figure 3-8* provides a detailed breakdown of the interviews conducted.

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. Interview protocols were developed for each major category of interviewee so that there would be consistency across interviewers. While the board members, association leadership, Design Team, central administration and external community members and funders were identified by their role in the district or the pilot, principals and teachers were drawn randomly from the population of pilot and control school principals and teachers. Parents were identified by various sources over the years of the pilot, including the Community Relations office, principals in pilot and control schools and parent-to-parent communications.

Objective-Focused Interviews

At the mid-point of the pilot a positive correlation was found between the quality of the teacher's objectives, as measured by the rubric, and student growth on the *ITBS* and the *CSAP*. This finding, and information from interviews and surveys,

FIG. 3-6

Distribution of Surveys

	1999-2000		2000-2001		2001-2002		2002-2003	
	Sent	Recd	Sent	Recd	Sent	Recd	Sent	Recd
Pilot School Surveys	420	349	400	362	617	330	604	395
Control School Surveys			660	243	855	330	850	278
Parent Surveys			1,200	122	2,580	104	3,602	357

FIG. 3-7

Distribution of Respondents

Survey Group	Demographic Characteristic	1999-2000	2000-2001	2001-2002	2002-2003
Pilot	Classroom Teacher	66.2%	64.5%	63.8%	60.6%
	Special Subject Teacher	10.0%	13.1%	12.4%	9.0%
	Special Education Teacher	8.0%	9.4%	11.1%	9.3%
	Special Services Provider	6.9%	8.9%	6.5%	10.4%
	School Administrator	3.2%	2.2%	4.0%	3.7%
Pilot	One Year in the District	11.0%	14.4%	15.6%	10.8%
	Two Years in the District	8.3%	9.7%	11.3%	14.1%
	Three Years in the District	6.4%	8.0%	4.0%	10.8%
	Four to 13 Years in the District	40.8%	36.0%	35.5%	38.3%
	14 or More Years in the District	33.4%	31.9%	33.6%	26.0%
Pilot	One Year in this School	22.2%	24.9%	28.1%	24.3%
	Two Years in this School	10.9%	14.7%	15.3%	17.0%
	Three Years in this School	11.6%	9.1%	8.5%	12.7%
	Four to 13 Years in this School	45.3%	42.4%	37.4%	36.7%
	14 or More Years in this School	10.0%	8.9%	10.7%	9.3%
Control	Classroom Teacher		57.4%	60.4%	59.6%
	Special Subject Teacher		10.7%	15.2%	8.5%
	Special Education Teacher		7.4%	9.8%	11.9%
	Special Services Provider		5.8%	4.4%	6.9%
	School Administrator		13.2%	8.9%	9.6%
	Other		5.4%	1.3%	3.5%
Control	One Year in the District		10.3%	11.4%	1.8%
	Two Years in the District		6.2%	10.2%	11.7%
	Three Years in the District		4.5%	7.4%	7.3%
	Four to 13 Years in the District		36.2%	34.8%	40.9%
	14 or More Years in the District		42.8%	36.3%	38.3%
Control	One Year in this School		20.7%	20.5%	6.0%
	Two Years in this School		13.3%	16.4%	21.8%
	Three Years in this School		8.7%	14.4%	10.7%
	Four to 13 Years in this School		48.1%	36.6%	48.8%
	14 or More Years in this School		9.1%	12.1%	12.3%

generated additional research questions around the process teachers used to develop objectives.

In order to address this issue, an objective-focused interview protocol was designed to ask a select number of teachers about the objective setting process. During the regular interview schedule in Spring 2002, 12 out of 64 teachers in seven out of 16 pilot schools were asked to describe their process for developing objectives for Pay for Performance. These teachers were a subset of the random sample of teachers who were chosen for interviews.

The interviews were analyzed for common themes and ideas regarding the process of developing objectives. While the number of teachers interviewed and schools represented are too small to generalize to the entire pilot population of teachers, the interviews provided insight into the objective development process as perceived and undertaken by 12 teachers, showing in particular, that teachers brought a range of thinking styles and pedagogical beliefs to the process. These results are discussed in Chapter IV.

FIG. 3-8

Distribution of Interviews by Role in District*

Role in the District	Number
Denver Classroom Teachers Association Leaders	20
Board Members (current and past)	31
Central Administration (including Superintendent)	49
Design Team Members	13
External Community Members	26
Other Site Staff	15
Parents	91
Principals	92
Teachers	278
Total	615

*This includes people who were interviewed in more than one year.

Qualitative Study

Based on the results of these objective-focused interviews and related findings from other interviews and surveys conducted in Spring 2002, deeper qualitative studies similar to case studies were conducted over a period of several months in the 2002-2003 school year. In trying to understand the relationships between setting an objective, meeting that objective, and improving student achievement on independent measures, CTAC staff designed a multi-method study based on the following proposition:

> *There is a positive relationship between the teacher objectives under PFP and changes in instructional preparation and classroom practices that research has shown to influence student achievement.*

A sample was selected by first identifying a set of indicators representing three categories: teacher demographics, student demographics and student achievement on the *CSAP:*

- Teachers Demographics: Percent w/Advance Degrees, Less than 3 Years at the School, Less than 3 Years of Experience, More than 10 Years Experience, Percent Hispanic Teachers, Percent White Teachers.
- Student Demographics: Percent Receiving Free/Reduced Lunch, English Language Learners, Student Mobility, Black Students, Hispanic Students, White Students.
- Student Achievement on the *CSAP:* Percent Advanced Category—Grade 3,4,5 Reading, Writing and Grade 5 Math; and Percent of Students in Unsatisfactory Category—Grade 3, 4, 5 Reading, Writing and Grade 5 Math.

Schools were ranked on these indicators and four schools were selected that best represented the schools in the Denver system. Four teachers were selected—three classroom and one specialist/special subject teacher from each of the four schools. Where possible, teachers who had been in the pilot for at least two years were selected. The four specialists/special subject teachers were selected to include a range of assignments: special education self-contained classroom, subject matter such as music or art, specialist such as a social worker/counselor.

Data collection involved three different visits by the research team: one visit to observe all 16 classrooms or workspaces for a full day, two visits to conduct partial-day observations and two 90-minute after-school focus groups at each school with the same 16 teachers. The visits were scheduled at two-month intervals: November, January, and March. A complete discussion of the findings of this study can be found in Chapter V.

G. Impact of School and Broader Institutional Factors

The pilot also exists in a broader district context. The institutional capacity to implement a major new initiative has been a factor in the success or failure of many other educational improvement initiatives. This capacity could also greatly affect the results of the pilot. For that reason, the study examined a range of institutional factors which might impact the pilot.

The decisions and actions of many participants within the institution can substantially influence the implementation of the pilot and its outcomes. This includes such pivotal groups as the Board of Education, the Association, the central administration, the Design Team and others.

The study examined policy and operational decisions, support structures, assignments, mid-course corrections and related interventions through the review of documentary data. The study also examined the perceptions of different constituencies—at the central and school levels—of these decisions and actions with yearly interviews.

The study further examined which efforts were perceived by various constituencies as supporting or impeding the progress of the pilot, the findings which have emerged, and the implications of those findings for the district in terms of the ability to implement major new initiatives. The final source of data concerning institutional factors came from participation in and observation of the processes.

Documents and Secondary Resources

Documentary data were collected from many sources including the following:

- Design Team: The Design Team provided 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, and other documents. In addition, representatives from CTAC attended monthly Design Team meetings and received copies of minutes of these meetings prepared by the Design Team.
- Administration, Board of Education and the Association: Documents requested from the district included board news and press releases, descriptive material on particular aspects of the pilot and internal newsletters and communications. The district also maintains considerable information concerning PFP and other topics on its website (www.dpskl2.org).
- Joint Task Force on Teacher Compensation and Leadership Team: CTAC's representatives attended monthly meetings of these two groups, received minutes of the meetings and other documents disseminated by the groups regarding the pilot and the proposed new compensation plan.
- 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, *Phi Delta Kappan,* the Business Roundtable, and others.

Interviews

These primary sources provide considerable insight into the actions of different entities and how they are perceived by people inside and outside the district. 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 over the evolution of the pilot. These interviews included discussions of various issues with board members, officials from both the district and DCTA, Design Team members, members of the corporate and philanthropic communities and teachers, principals and parents.

CTAC Participation in and Observation of Processes

Another source of information regarding the impact of the broader institutional factors involved the participation of and observation by representatives of CTAC at meetings of the Design Team, the Joint Task Force on Teacher Compensation, the Leadership Team, and the Communications Group in addition to individual meetings with the superintendent, the Board's liaison for the pilot, and Association leadership. CTAC representatives participated in regular monthly meetings with these groups and presented reports on aspects of the research study to various constituent groups.

A project or pilot can only be successful if it can be implemented. The national experience in school reform has repeatedly demonstrated the widely varied impact that different implementation strategies and approaches have had on results, even when the programs were similar. Accordingly, the study has paid close attention to issues of implementation—both how things are done and how they might 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 throughout this report.

H. Summary

The study of Pay for Performance in Denver was designed to examine the impact of linking student achievement to teacher compensation. Moreover, the study also examines the school level and broader institutional factors which may have influenced the implementation of the pilot.

Both quantitative and qualitative data were collected and analyzed over a four-year period: student achievement results; school, teacher, and student factors; artifacts; participant surveys; participant interviews; and observations. Several types of quantitative analyses have been conducted, including: two-stage hierarchical linear modeling; individual growth modeling; simple linear regression analysis; and rubric-based analyses.

Other steps have been taken to ensure the rigor of the study and to probe the findings from some of the data sets. These include the identification of a group of control schools whose student performance could be compared to that of the pilot schools and the development of deeper qualitative studies in order to probe specific findings.

These are areas that Denver will continue to address as PFP is potentially implemented full-scale in the district.

CHAPTER IV

Objectives: The Nexus

A. Introduction

The heart of the Pay for Performance pilot is the teacher objective setting process. Pilot school teachers individually developed two yearlong instructional objectives for each of the four years of the pilot, using the following process: (1) review the available baseline achievement data on their current year students; (2) write two objectives for the identified population(s); (3) select a measure for each objective; (4) establish expected gain or growth targets for the students in the class; and (5) confer with the building principal for approval. At the end of the school year, the teacher presented evidence that one or both objectives had been met, and if the principal concurred, the teacher was compensated commensurately. In actual practice, objective setting for the pilot also called for teachers to write a rationale and teaching strategies and, over the course of the pilot, has required the use of various written formats. Through this process, instructional objectives became the nexus between teacher performance and student performance that results in additional compensation.

Instructional objectives that identify what teachers will teach and what students will learn have a long-established currency in educational settings. They are the hallmark of instructional planning for the year, the unit, and the lesson. Often, such objectives are written for teachers and can be found in curriculum guides, on-line data banks, and textbook publisher materials; however, teachers are also called upon to write objectives (or goals or outcomes) for many purposes in their yearly work, and objective writing as the beginning of instructional planning is a topic in most teacher training programs. However, writing objectives for compensation requires better information and greater precision than is customarily associated with planning objectives.

Basing additional compensation on the results of teacher-developed instructional objectives is both the *inspiration* in the design of the Pay for Performance pilot and the *agency of many of the dilemmas* of its implementation. The inspiration is in the appropriation of an existing district practice, one of writing annual goals or objectives for the teacher appraisal process, in order to house a potentially controversial reform. Developing two or three goals or objectives and submitting them to the building principal is a familiar routine in Denver schools and, more importantly, a practice where teacher autonomy is well established. Educators who implement reforms recognize the importance of moving participants from the familiar to the new. Similarly, teacher leaders who negotiate contracts understand the significance of obtaining and maintaining teacher autonomy in district mandates to the highest degree possible. Teacher objectives, as the base component of the pilot, provided the district and the Association with a familiar launch point to test a new approach to compensation, and then, as a teacher-developed product, objectives contributed a significant level of teacher autonomy to a high stakes reform.

On the other hand, teacher objectives, which are the intended drivers of the pilot, have, paradoxically, been the agency of many of the dilemmas in the implementation of the pilot, both creating new issues to be resolved and encountering barriers within the system. For example, persuading schools and teachers to join the pilot with promises of earning additional compensation for "doing what you already do" introduced elements of past practices into the implementation of a new initiative. This marketing feature of the pilot may have also set teacher participants down a determined path of "not changing what I do," an unintended consequence that is explored more extensively in Chapter V. Secondly, developing teacher objectives that are "data driven, credible, and fair" (Design Team Project Plan, 2000) has been limited by systemic barriers, the most daunting of which was the lack of aligned and consistently administered assessments. Alignment and assessment issues are discussed in Chapter VIII.

As this chapter and the subsequent one will demonstrate, setting objectives that lead to improved student achievement and increased compensation requires a higher level of science on the part of teachers, principals, and district leaders than the routine setting of goals or objectives where there may be little or no accountability for the outcome. Changing the customary and less scientific mode of writing and assessing objectives into a more reliable process became the ongoing work of the Design Team. Consequently, improvement in the setting of the objectives over the course of the pilot has resulted in increased numbers of objectives that meet the quality criteria, which are explained and discussed later in this chapter. Secondly, higher average student achievement on independent measures is associated with higher quality objectives and with the number of objectives met.

This chapter addresses the topics of (1) the complexity of implementing objectives; (2) the quality criteria for the objectives, including methodology for and the results of the holistic scoring; and (3) the results of comparisons of objective data to other available data sets.

B. Unexpected Complexity and Barriers to Implementation

The Complexity of Implementation

Interview and observation data from the study show that many teachers in Denver consider the crafting of objectives a long-standing and routine part of their work, something that they have always done. Also, responses to interview questions about objective setting show that over the course of the pilot, accountability for reaching objectives has entered teacher and principal discourse—both positively and negatively. For some, the objective setting process is a variation on "business as usual"—for others, it has increased critical thought and reflection about teaching and learning. The following excerpts from teacher interviews demonstrate this point:

> "Objective setting [is] always the same. The only difference is the structure and the reward. And I look at them more than before."
> –Pilot teacher

> "I have always set objectives and had rubrics to see what I wanted to achieve. But for the two PFP goals I am more specific."
> –Pilot teacher

"I have learned about the importance of setting specific goals using assessment data." –Pilot principal

"Objectives now determine yes or no to receiving compensation. It makes you think about the objective and work toward it. I don't have time to chase objectives down. Teaching is already a busy profession. PFP is not inherent, at the end of the day, to whether or not I have achieved my goals. I hope that by teaching a strong curriculum and providing support for testing it will work."–Pilot teacher

"Last year, I had a very sincere, heartfelt objective for students to learn about syntax. What I felt after struggling with it for that year is that I really learned a lot about my students, what worked and what didn't. Even though I didn't make the objective (by a very small bit) it was a good experience. It was what we learned through the process and what happened during the year that mattered, not the pay part." –Pilot teacher

"I have always written objectives. We have had to become more specific in our goals, percentages, although we did that before. We are also looking at goals that are reachable. So that is why we look at scores carefully and then we work to meet those goals. In the way we write goals, there is no change in content. We just make sure that it is measurable and very specific. No major change."–Pilot teacher

These remarks also show teachers grappling with the differences between the old process and the new one. There is recognition, if not clarity, about the importance of specificity, measurement, and accountability in the new process; yet, these teachers clearly still consider objectives their domain and within their mastery.

Nonetheless, objective setting for PFP turned out to be unexpectedly complex. In *Pathway to Results,* the mid-point report, pilot participants from board members to teacher leaders to classroom teachers and principals report on their surprise that objective writing could be so complex and create so many dilemmas. Two board member comments, for example, early in the implementation show this surprise:

"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."–Board member

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

And in interviews during the last year of the pilot (Spring 2003), teachers and principals were still pondering the issues and challenges of using objectives as the basis of bonus pay.

"We need uniform procedures for objectives. The timeline didn't make sense. We started the school year in mid-August, goals written by mid-October, reviewed by November or December. It should have been done earlier in the year. I didn't get the 'OK' on my objectives until December."–Pilot teacher

"We have to find a way to distinguish between school politics and the pecking order in schools and actual teacher performance. Right now those are meshed together. Right now the leadership has too much decision-making power and that causes anxiety."–Pilot teacher

"Teacher graded assessments leave opportunities to manage the outcome."–Pilot teacher

But a major change had taken place from the earlier interviews (Spring 2000) to the more recent (Spring 2003). Fewer participants were saying, "It just won't work." More were identifying weak points and suggesting changes and repairs to the process in order to make it work more effectively.

At the end of four years, there is a greater appreciation on the part of teachers, administrators, and Design Team members for the complexity of setting objectives for compensation purposes. Many teacher objectives at the outset of the pilot were focused on improving student performance on an assessment (i.e., "70% of my students will gain one year or more in reading on the *Iowa Test of Basic Skills*") rather than focusing on learning content. Secondly, student achievement measures were not aligned to state and district standards. Finally, there had been inadequate professional development for both principals and teachers on the craft of setting

objectives and aligning instructional practices to them prior to the pilot. With only brief sessions on objective setting in the first year (1990-2000) and without district direction and support on connections between objectives and teaching, the customary way of writing goals and objectives prevailed in the early implementation of the pilot. As a result of these issues, concerns around the measurement, consistency, and fairness of the objectives emerged among participants.

Over the course of the pilot, more technical assistance and training was provided to teachers and principals by Design Team members in order to improve the quality of objectives, including, "how to" descriptions and rubrics that guided the development of objectives for classroom teachers, special subject teachers, and specialists. Also, a database of student assessment information called Online Assessment Score Information System (OASIS) was developed for pilot teacher use in May 2001. Improvement in the quality of the objectives is documented later in this chapter. Also, inroads have been made into many of the systemic barriers to implementation of a pay for performance system based on objectives, though key systemic processes still need to be addressed.

Pilot Year One 1999-2000
A small staff trained the initial group of elementary pilot schools in setting objectives for PFP. The use of baseline or pre-test data was emphasized because of the need to measure student growth to receive bonus pay. Many of the examples of objectives provided were in the style of pre-existing district practice, improvement on assessments. Some follow-up was provided to teachers.

Pilot Year Two 2000-2001
The Design Team introduced a worksheet and heuristic or template for teachers to complete with the following categories: objective, population, assessment, baseline data, rationale, teaching strategy, and evidence.

Pilot Year Three 2001-2002
The Pay for Performance summer training presented the key tasks of writing an objective integrated with planning documents for the teacher's use. These included using the Denver Standards and Curriculum Matrices; how to analyze assessment trends in order to assist in the objective setting process; and developing a body of evidence among others. The Design Team and the district also introduced the use of OASIS for teachers to find the assessment history of their students online along with a web-based system where teachers input their own objectives into the new format. The categories of the new format are similar to previous years except that the "objective" category is not included, reflecting an expectation that the components of the heuristic will add up to the total objective. The web-based system improved the quality of the objective information and decreased technical errors.

Pilot Year Four 2002-2003
Work with groups of teachers by the Design Team resulted in the development of analytical rubrics for classroom teachers and special subject teachers and a checklist for specialized service providers. Exemplary objectives were developed for elementary and secondary teachers based on the rubric. A key change on the rubric—adding a Learning Content category—as well as a similar change in the Web-Based Objectives format and examples of objectives using the new category led to higher levels of scores on the pilot research rubric (discussed later in the chapter). Finally, the Design Team initiated and validated a rubric-based evaluation system of its own.

Systemic Barriers to Implementation

Barriers to quality objective setting existed not only at the teacher and school level but also throughout the system. At the outset, there was a lack of alignment between district content standards and assessments, and hundreds of assessments, some teacher-made and almost all teacher-administered and scored, were in play. District direction on the appropriate and consistent use of designated assessments for the district standards was absent, incomplete, or implemented unevenly. Administration of the norm-referenced *Iowa Test of Basic Skills* was permissive for many schools, and the new Colorado state test *(Colorado Student Assessment Program)* was just emerging by grade levels. Since no performance standards or annual expected gain for students had been established for teacher use, teacher expectations for student growth varied from school to school and from teacher to teacher in the same school.

Besides assessment, there were other unaddressed systemic issues that complicated the implementation of objectives: (1) reliable integrated student and teacher data; (2) a fully developed plan that integrated the PFP elements with district and school educational plans; and (3) focused professional development for teachers and principals in pilot schools. Addressing or troubleshooting as many systemic barriers as possible in order to improve the quality, rigor, and consistency of the objectives constituted a large block of the Design Team's implementation work.

FIG. 4-1

Traits or Criteria for Quality Educational Objectives

Trait One: 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 Two: Completeness
A complete expression of an educational objective includes: the student population to be taught; the objective with learning content; the assessment; the strategy or strategies used by the teacher to address the content; the rationale for selecting the objective; baseline data that show prior knowledge and/or skills; and finally, the evidence that persuades the teacher that the objective has or has not been met.

Trait Three: 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 Four: 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/herself.

The Design Team Implementation of PFP Objectives

The newly appointed Design Team members started up the implementation of the objectives element of the pilot in the fall of 1999 almost synchronously with recruiting schools into the pilot and with little time to plan. Over the course of the four years, the team has refined the "how to" information for teachers and improved the beginning-of-school training sessions. In the fall of 2002, teachers received a highly professional handbook to assist their objective setting and instructional planning process. In the last year of the pilot, a focus group of teachers remarked on the quality of the training materials, wishing that they had been available in the early years of the pilot.

Learning what teachers needed in order to develop objectives for compensation is a key outcome of the pilot. Following the progression of annual training materials tells a story of continuous research-driven improvement in the work of the Design Team with teachers that resulted in yearly improvements in the quality of the objectives.

A significant part of the learning of the pilot can be seen in the progression of annual training sessions on the writing of objectives. As a previously quoted interviewee remarked: "Who could have thought it would be so hard to write two objectives?" An existing practice of setting objectives in Denver was full of good intent and communicated well enough within the school, but it was not adequate for use in a compensation program. The Design Team continues to refine the process so that teachers have a stronger notion of what is involved in developing a measurable objective, particularly, the use of baseline data and learning content, and so that they appreciate the potential of greater focus and more precision in measurement.

C. Quality of Teacher Objectives

A major charge of the pilot study was to determine the quality and impact of the objectives. Just as the implementation of the objectives element of the pilot presented the Design Team, teachers, and principals with a complex set of issues, developing a process to assess the quality and impact of the objectives presented the research study team with some methodological challenges.

Not finding an accepted evaluation tool for determining the quality traits of an instructional objective in the research literature, a panel of

FIG. 4-2

Levels of Performance

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 inconsistent 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.

educators examined (1) the literature and guides for teacher planning in the ERIC system: (2) Denver's scope and sequence for K-12; and (3) the heuristic template provided to Denver teachers for writing an objective.[1] No method or style of objective writing emerged in the literature as more effective than another in getting results. The behavioral objectives in vogue in the 1960s and 1970s that included the "elements of performance, conditions, and criterion"[2] have not been associated with significant gain.[3] However, there is a stream of research to indicate that teacher lesson planning is associated with student gains, and objectives are the accepted first step of an effective planning process.[4] There is some relatively recent research, however, which indicates that overly specific or narrow goals are negatively correlated with student gain.[5]

While there is not a research-based method or even a clearly preferred model for writing instructional objectives, a review of models found in the lesson planning literature indicates that instructional planning includes: (1) what will be taught (standards, concepts, skills, etc.); (2) how students will demonstrate learning (assessments, products, performances, etc.); and (3) teaching strategies. So it was from practitioner planning literature that the key traits of quality educational objectives were derived for the study.

Methodology

In order to carry out the evaluation of objectives, CTAC developed a rubric for the holistic rating of objectives. The first stage of developing the rubric was to identify the traits of quality educational objectives. The categories of traits derived from the review of examples in the literature and the heuristic format provided to pilot participants include: (1) *learning content,* what the teacher will teach and the student will learn; (2) *completeness,* the use of seven elements from the heuristic format provided teachers by the Design Team; (3) *cohesion,* the logic and unity among the elements; and (4) *expectations,* the expected level of student growth anticipated by the teacher. *Figure 4-1* describes these criteria.

The second stage of developing a rubric was the development of levels of performance. A ranking of second year objectives contributed to the final assignment of the performance levels of *Excellent, Acceptable, Needs Improvement,* and *Too Little to Evaluate.* The decision to use a four-point scale over a six-point scale was based on the observation that there was not enough substance in the objectives to discriminate among six levels and, of course, on the need for expediency in processing the large number of objectives each year. The performance levels are shown in *Figure 4-2.*

The final stage of rubric development integrates the four traits or criteria into descriptors for each of the four performance levels. The rubric is shown in *Figure 4-3.*

A panel of readers with teaching and curriculum administration experience and expertise rated all of the objectives based on the rubric. Discrepancies in ratings among readers were resolved through a second, and if needed, third rubric-based reading and discussion.

For the purposes of comparisons over the life of the pilot and the identification of trends, it was important to maintain the same rubric over the life of the pilot. As discussed earlier, the support provided annually by the Design Team resulted in different heuristic devices and formats provided to the pilot teachers for each of the four years of the pilot. For this reason, the readers of the objectives re-anchored each year. However, the rubric remained robust through the changes.

Results of the Rubric-Based Evaluation, 1999-2003

The results of the rubric-based evaluation for each of the four years are shown in *Figure 4-4.* The majority of the objectives for years one and two of the pilot fall into the second performance level, *Needs Improvement;* in the second year of the pilot, the percentage of objectives in the level 2 category decreased substantially from 61% to 54% and the percentage designated at level 4 grew by eight percentage points to almost 9%. Level 3 remained relatively constant in the first two years.

As discussed extensively in the mid-point report, most teachers scored lower than might have been anticipated in the first two years of the pilot, an outcome largely attributable to the fact that learning content, one of the rubric traits, was

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 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 the learning content, showing a lack of understanding about what is expected or confusing the elements (stating the objective as an assessment goal rather than a learning goal). Expectations for students may be 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 to be reliably assessed.

missing from most objectives. Where it was present, it was often of a general nature (i.e., reading, mathematics). Following an existing practice in the schools, many teachers wrote their objectives as improvements of assessment performance rather than of learning the content. An example of this type of assessment-focused objective is as follows: "75% of the identified students will show a growth of one year or more on the *Developmental Reading Assessment (DRA)/Qualitative Reading Inventory (QRI)*." Another influence on the use of assessment-focused objectives was the designation of approaches in the original pilot design, two of which were intended to examine the use of specific types of assessments.

At the time of the mid-point report, the research team assessed the significance of setting objectives in this manner and determined it important to keep the content trait as part of the rubric because it is the content that communicates what is being taught. Identifying the content to be taught also reduces the likelihood that the assessment will be perceived as the content (teaching to the test); and finally, it increases the likelihood that teacher reflection and planning will focus on content alignment and attainment, factors likely to improve student achievement. This topic is discussed extensively in the mid-point report (pp. 32-36).

The expectation trait of the rubric was also a pitfall in the first year of the pilot as teachers sought reasonable growth targets for their students, one that is challenging but reachable. In the second year, expectations grew, and by the third and forth year of the pilot, a typical growth target within a teacher objective had settled on 75% of the students who were present 85% of the year.

The third year of objective scores (2001-2002) show additional increases in the percentage of level 4 scores (to 13%) and level 3 scores (to 34%) with the majority of objectives (52%) remaining at level 2 on the performance scale. During this year, the Design Team and the district introduced the use of OASIS, where teachers could access prior student assessment data and the Web-Based Objectives software for inputting their objectives. In the first two years of the study the objectives were transferred into electronic format for analytical use. With the introduction of the Web-Based Objectives system, the percentage of objectives that were incomplete or contained errors declined. It is possible that, in earlier years, some of the objectives that could not be rated were missing information due to transcription errors.

In the fourth year (2002-2003) of the pilot, objective scores improved dramatically with the percentage of level 4 scores more than doubling (to 28%), the level 3 scores increasing (to 44%), and concomitantly, the level 2 scores decreasing

FIG. 4-4

Summary of Rubric Levels, 1999-2003

Year	Rubric Level	First Objective	Second Objective	Both Objectives	Percent
1999-2000	4	1	5	6	0.9
	3	72	93	165	24.1
	2	199	220	419	61.3
	1	51	1	52	7.6
	Unrated	19	23	42	6.1
	Total	342	342	684	100.0
2000-2001	4	32	38	70	8.9
	3	82	96	178	22.6
	2	223	203	426	54.1
	1	54	52	106	13.5
	Unrated	3	5	8	1.0
	Total	394	394	788	100.0
2001-2002	4	80	89	169	13.2
	3	202	234	436	34.1
	2	355	307	662	51.7
	1	3	8	11	0.9
	Unrated		2	2	0.2
	Total	640	640	1280	100.0
2002-2003	4	179	174	353	28.0
	3	281	276	557	44.2
	2	168	171	339	26.9
	1	2	2	4	0.3
	Unrated		7	7	0.6
	Total	630	630	1260	100.0

by one half. A change in the analytical rubric developed by the Design Team, along with a change in the structure of the Web-Based Objectives software format, prompted most teachers to include the content to be taught in their written objectives. More clearly articulated content in objectives accounts for most of the improvement in scores. However, the increased use of content statements also reveals that teachers often have difficulty in connecting all of the pieces listed in the format into a coherent whole (i.e., measuring what students know and what they will learn; holding high expectations for students; and being thoughtful and complete in writing their objectives) so that merely adding content did not necessarily create a level 4 objective. In fact, the requirement to respond to the new learning content category may have been confusing. For example, in the learning content category, teachers sometimes listed teaching strategies (how, not what) or the content topics for the entire year's curriculum, or they reversed the rationale and content categories on the format, affecting the cohesiveness trait of the rubric. These findings

FIG. 4-5

Numbers and Percentages of Objective Rubric Levels by School by Year, 1999-2003

School	Year	Total 1s	Total 2s	Total 3s	Total 4s	Total Scores	% 1s	% 2s	% 3s	% 4s
Centennial	2000	3	47	19	1	70	4.3	67.1	27.1	1.4
	2001	8	48	19	1	76	10.5	63.2	25.0	1.3
	2002		40	30	6	76		52.6	39.5	7.9
	2003		14	49	13	76		18.4	64.5	17.1
Colfax	2000		43	6	1	50		86.0	12.0	2.0
	2001	2	41	3	6	52	3.8	78.8	5.8	11.5
	2002		45	5	4	54		83.3	9.3	7.4
	2003		6	35	11	52		11.5	67.3	21.2
Columbian	2000	1	38	7		46	2.2	82.6	15.2	
	2001		23	5	4	32		71.9	15.6	12.5
	2002		24	13	1	38		63.2	34.2	2.6
	2003		10	27	5	42		23.8	64.3	11.9
Cory	2000		5	41		46		10.9	89.1	
	2001		21	14	15	50		42.0	28.0	30.0
	2002		21	18	11	50		42.0	36.0	22.0
	2003		6	33	17	56		10.7	58.9	30.4
Edison	2000	3	52	3		58	5.2	89.7	5.2	
	2001	3	39	21	1	64	4.7	60.9	32.8	1.6
	2002	1	33	24	6	64	1.6	51.6	37.5	9.4
	2003		21	19	20	60		35.0	31.7	33.3
Ellis	2000	35	35			70	50.0	50.0		
	2001		24	38	6	68		35.3	55.9	8.8
	2002		39	22	9	70		55.7	31.4	12.9
	2003		15	33	24	72		20.8	45.8	33.3
Fairview	2000		42	12		54		77.8	22.2	
	2001	12	26	12	5	55	21.8	47.3	21.8	9.1
	2002	3	42	7	4	56	5.4	75.0	12.5	7.1
	2003		26	22	14	62		41.9	35.5	22.6
Mitchell	2000		33	33		66		50.0	50.0	
	2001	7	33	13	6	59	11.9	55.9	22.0	10.2
	2002		39	27	6	72		54.2	37.5	8.3
	2003		16	36	22	74		21.6	48.6	29.7

FIG. 4-5 CONTINUED

Numbers and Percentages of Objective Rubric Levels by School by Year, 1999-2003

School	Year	Total 1s	Total 2s	Total 3s	Total 4s	Total Scores	% 1s	% 2s	% 3s	% 4s
Oakland	2000		59	9		68		86.8	13.2	
	2001	61	9			70	87.1	12.9		
	2002		49	14	15	78		62.8	17.9	19.2
	2003		23	44	12	79		29.1	55.7	15.2
Philips	2002		38	10	10	58		65.5	17.2	17.2
	2003		26	16	12	54		48.1	29.6	22.2
Smith	2000	8	27	3		38	21.1	71.1	7.9	
	2001		61	5		66		92.4	7.6	
Southmoor	2000			19	1	20			95.0	5.0
	2001		10	19	5	34		29.4	55.9	14.7
	2002		8	30	2	40		20.0	75.0	5.0
	2003			17	27	44			38.6	61.4
Traylor	2000	2	38	13	3	56	3.6	67.9	23.2	5.4
	2001	2	37	6	15	60	3.3	61.7	10.0	25.0
	2002		42	15	7	64		65.6	23.4	10.9
	2003		3	45	14	62		4.8	72.6	22.6
Horace Mann MS	2001	11	54	23	6	94	11.7	57.4	24.5	6.4
	2002		42	49	17	108		38.9	45.4	15.7
	2003	1	33	35	23	92	1.1	35.9	38.0	25.0
Lake MS	2002		71	48	13	132		53.8	36.4	9.8
	2003		47	35	38	120		39.2	29.2	31.7
Manual HS	2002	5	62	68	33	168	3.0	36.9	40.5	19.6
Arts & Culture	2003	3	31	17	12	63	4.8	49.2	27.0	19.0
Leadership	2003		17	20	19	56		30.4	35.7	33.9
Millennium	2003		13	21	14	48		27.1	43.8	29.2
Thomas Jefferson HS	2002	2	67	56	25	150	1.3	44.7	37.3	16.7
	2003		32	53	56	141		22.7	37.6	39.7
Total	2000	52	419	165	6	642	8.1	65.3	25.7	1.0
	2001	106	426	178	70	780	13.5	54.6	22.8	9.0
	2002	11	662	436	169	1278	0.9	51.8	34.1	13.2
	2003	4	339	557	353	1253	0.3	27.1	44.4	28.2

suggest that some teachers near the end of the pilot were continuing to struggle with objective setting.

The 2002-2003 objective format was the fourth one in as many years that pilot teachers used in order to write their objectives. Each new format template represented an improvement over the previous year's format but a new set of challenges for teachers, principals, and researchers. Several of the teachers in focus groups recognized the 2002-2003 Design Team rubric and support materials as superior tools and wished that they had been available in the first years of the pilot, but for other teachers in the study, it was just another new form and a bit more aggravation when they already had their process down.

As in the previous year, the use of the Web-Based Objectives computer program enhanced the year's objective format. The teacher had to fill in each of the categories to complete the process, reducing the chances that a rubric level would be based on a partial document. Where an objective was incomplete or not available at the time of the rubric analysis, it was not rated and is shown as "unrated" in *Figure 4-4*. These numbers were small and are not included in subsequent figures. Annual changes in the directions and formats for the objectives made each new set of objectives a challenge for evaluators, both in attempting to maintain a consistent application of the research rubric for the reliability of the study and in overcoming the different technical problems that each format presented.

D. Research Questions, Data Sources and Findings Related to Objectives

By developing and applying the rubric, the research team began its study of pilot teacher objectives, answering the first of the research questions about this element of the pilot design.

1. What are the traits of a quality objective and how are they best described?

Having developed a rubric with which to evaluate the objectives, the next step was to apply the rubric to the objectives written by pilot teachers. Did pilot teachers write quality objectives? Is there a relationship between teacher characteristics and the quality of objectives? Did the writing of objectives translate into higher student achievement? These issues are expanded upon in the following research questions:

2. What are the rubric levels of objectives written by teachers? Is there a relationship between the quality of the objective written by the teacher and the teacher's participation in the Teacher-in-Residence program, years of experience in the Denver schools, educational background, and years of participation in the pilot?
3. Is there a relationship between the quality of the objective written by the teacher and student achievement as measured on an independent, standardized test that measures general growth?
4. Is there a relationship between whether a teacher meets his or her objectives by the measures or parameters he or she has set and the teacher's participation in the Teacher-in-Residence program, years of experience in the Denver schools, educational background, and years of participation in the pilot?
5. Is there a relationship between whether a teacher meets his or her objectives by the measures or parameters he or she has set and student achievement on an independent, standardized measure of general growth?
6. Is there a relationship between the quality of a teacher's objective and the process he or she describes for writing, teaching to, and assessing that objective? (See Chapter V.)
7. Is there a relationship between teacher objectives and school improvement plan goals and objectives?
8. Do objectives written in the pilot schools differ substantially from those of teachers in control schools?

The rubric ratings are compared with other data sets in an effort to answer the research questions outlined in the chart. The data sets include: (1) four years of rubric levels for two objectives for all teachers in the pilot; (2) four years of achievement data; (3) four years of met/not met data—the

FIG. 4-6

Research Questions about Objectives and Data Sources

Questions	Data Sources			
	2000	2001	2002	2003
1. What are the traits of a quality objective and how are they best described?	Rubric Objectives	Rubric Objectives	Rubric Objectives	Rubric Objectives
2. What are the rubric levels of the objectives written by pilot teachers? Is there a relationship between rubric level and teacher characteristics?	Human Resource Files Rubric Level	Human Resource Files Rubric Level	Human Resource Files Rubric Level	Human Resource Files Rubric Level
3. Is there a relationship between the quality of the objective written by the teacher and student achievement on an independent, standardized measure, which measures general growth?	ITBS CSAP Rubric Level	ITBS CSAP Rubric Level	ITBS CSAP Rubric Level	ITBS CSAP Rubric Level
4. Is there a relationship between whether a teacher meets his or her objectives by the measures or parameters he or she has set and teacher characteristics?	Human Resource Files Met/Not Met Results	Human Resource Files Met/Not Met Results	Human Resource Files Met/Not Met Results	Human Resource Files Met/Not Met Results
5. Is there a relationship between whether a teacher meets his or her objectives by the measures or parameters he or she has set and student achievement on an independent, standardized measure, which measures general growth?	Met/Not Met Results ITBS CSAP	Met/Not Met Results ITBS CSAP	Met/Not Met Results ITBS CSAP	Met/Not Met Results ITBS CSAP
6. Is there a relationship between the quality of a teacher's objective and the process he or she describes for writing, teaching to, and assessing that objective?	General Interviews	General Interviews	Process-focused Interviews	General Interviews
7. Is there a relationship between teacher objectives and school improvement plan goals and objectives?		School Plans		School Plans
8. Do objectives written in the pilot schools differ substantially from those of teachers in control schools?		Control School Teacher Goals		Control School Teacher Goals

numbers and percentages of teachers meeting their objectives; (4) four years of survey data; (5) four years of interviews; (6) samplings of other artifact data; and (7) specialized interviews, focus groups, and observations; and (8) four years of teacher characteristics from the DPS Human Resource files. *Figure 4-6* shows the questions and data sources used.

The research questions are primarily addressed in the remainder of Chapter IV. Question 6 is explored in length in Chapter V. The relationship between objectives and school improvement plans, articulated in question 7, is at root an issue of instructional and organizational alignment and, therefore, is explored in Chapter VIII.

Analyses of Teacher Objective Data

In exploring the connection between student achievement and objectives, the study links the teacher who wrote the objective to the students

he or she taught and ultimately to the achievement scores of those students. However, many teachers do not have easily defined classes or caseloads of students. The analyses that follow concentrate on the subset of objectives written by classroom teachers. *Figure 4-7* describes the objectives written by classroom teachers who could be linked to specific students.

Objective Quality and Teacher Characteristics

Aggregated over the entire four years of the pilot (see *Figure 4-8*), teacher educational level is not related to the rubric level of classroom teacher objectives. Teachers-in-Residence (TIRs) are both new to the teaching profession and lack an academic background in education; yet, the distribution of rubric levels for TIRs does not differ significantly from that of other teachers.

The relationship between a classroom teacher's length of experience in the Denver schools and rubric level is not significant when years of experience are categorized in four groups; however, it is significant when we focus on first year teachers. Twenty percent of first year teachers, as opposed to 6% of more experienced teachers, wrote a level 1 objective. First year teachers were also more likely to write level 2 objectives and less likely to score level 3 or 4 on the rubric. This finding has implications—objective setting skills need to be more explicitly addressed in the orientation of teachers to the Denver school system, and extra guidance from principals or mentoring teachers may also be beneficial.

Encouragingly, there is a significant increase in the rubric level of objectives as the number of years a classroom teacher participated in the pilot increases. This finding mirrors that seen for all objectives as referenced in *Figure 4-4*.

FIG. 4-7

Objectives Written by Classroom Teacher Characteristics by Year

Characteristic	1999-2000	2000-2001	2001-2002	2002-2003
	Percent (N)	Percent (N)	Percent (N)	Percent (N)
Rubric Level 4	0	4% (16)	6% (30)	21% (103)
Rubric Level 3	19% (62)	19% (71)	25% (126)	50% (240)
Rubric Level 2	73% (242)	57% (208)	67% (335)	29% (138)
Rubric Level 1	9% (29)	20% (72)	1% (6)	0
Objective Met	91% (321)	92% (335)	90% (442)	92% (449)
Teacher-in-Residence	1% (2)	5% (19)	10% (48)	13% (63)
Bachelor's Degree	49% (140)	53% (187)	58% (288)	63% (299)
Master's Degree	51% (148)	47% (166)	42% (208)	37%(175)
Doctorate	0	0.3% (1)	0.2% (1)	1% (4)
0 to 3 Years Experience	23% (68)	25% (89)	28% (134)	22% (80)
4 to 10 Years Experience	20% (60)	14% (51)	15% (73)	17% (62)
11 to 14 Years Experience	28% (84)	31% (111)	25% (118)	28% (102)
15 or more Years Experience	29% (86)	30% (108)	32% (156)	34% (126)
First Year Teachers	5% (14)	21% (75)	0	0
1 Year of Pilot Participation	100% (352)	26% (90)	25% (86)	25% (88)
2 Years of Pilot Participation		74% (260)	26% (92)	19% (66)
3 Years of Pilot Participation			49% (172)	20% (70)
4 Years of Pilot Participation				36% (124)

FIG. 4-8

Classroom Teacher Objectives—Rubric Level by Teacher Characteristics, 1999-2003

Teacher Characteristic	Rubric Level 1	Rubric Level 2	Rubric Level 3	Rubric Level 4
Teacher-in-Residence	Percent (N)	Percent (N)	Percent (N)	Percent (N)
No	6% (96)	55% (858)	30% (459)	9% (135)
Yes	8% (11)	50% (65)	31% (40)	11% (14)
$\chi^2 = 2.2$, $p(\chi^2=0) = 0.528$				
Educational Degree				
Bachelor's	6% (49)	56% (500)	30% (266)	9% (76)
Master's	7% (46)	53% (363)	30% (209)	10% (70)
Doctorate		83% (5)	17% (1)	
$\chi^2 = 5.1$, $p(\chi^2=0) = 0.536$				
Years of Experience in DPS				
0 to 3	8% (30)	55% (205)	28% (105)	8% (30)
4 to 10	5% (12)	53% (127)	32% (77)	10% (23)
11 to 14	7% (28)	54% (222)	31% (129)	8% (31)
15 or more	7% (30)	62% (284)	25% (112)	7% (31)
$\chi^2 = 12.2$, $p(\chi^2=0) = 0.201$				
First Year Teachers				
First Year	20% (17)	61% (53)	16% (14)	3% (3)
Subsequent Years	6% (82)	55% (784)	30% (424)	9% (129)
$\chi^2 = 32.4$, $p(\chi^2=0) = 0.001$				
Years of Pilot Participation				
1	8% (49)	63% (376)	25% (151)	3% (19)
2	12% (51)	57% (233)	24% (97)	7% (29)
3	1% (2)	57% (138)	32% (76)	10% (25)
4		22% (27)	59% (73)	19% (24)
$\chi^2 = 157.1$, $p(\chi^2=0) = 0.001$				

Rubric Levels of Objectives and Student Achievement

The third research question explores the relationship between student achievement and the quality of classroom teacher objectives. Mean achievement scores were estimated for elementary and middle school students by the maximum rubric level of their teacher, adjusting for student and school characteristics. Mean scores were estimated separately for each pilot high school as well, adjusting for student characteristics. For the secondary analyses, one language arts and one math teacher were selected at random for each student. For the most part, elementary school students spend the majority of the school day with one teacher, however secondary students may have a number of teachers who could be expected to impact the students' standardized test scores. The secondary school analysis is biased toward finding no relationship between achievement and rubric level, since the students for whom

we randomly chose a teacher with a rubric level of 1 may also have one or more teachers in other related classes with higher rubric levels. The full description of the student achievement analyses is presented in Chapter VI, and a summary of the findings is presented here in *Figure 4-9*. There were years in which none of the students who took the *ITBS* or *CSAP* exams had classroom teachers with a rubric level of 1, resulting in no estimates for rubric level 1 on those tests.

Elementary Schools

At the elementary school level, there is evidence that mean student achievement NCE scores increase as rubric level increases:

- On the *ITBS* Reading test, students of teachers with rubric levels of 2, 3, and 4 have mean NCE scores that are significantly higher than students of teachers with a level of 1.
- On the *CSAP* Writing test average scores increase as the rubric level of the students' teacher increases, however the differences between levels is not statistically significant.
- On the *ITBS* Language and *CSAP* Math tests, the average scores of students with rubric level 4 teachers are significantly higher than the other levels.
- Mean scores for the *ITBS* Math test are higher at rubric levels 2 and 3 than either level 1 or level 4.
- Mean scores for the *CSAP* Reading test are approximately equal across rubric levels.

Overall, on four out of six tests, there is a positive relationship between achievement and rubric level. On one test, the results are mixed, and on one test, there is no difference. This result is strong evidence that higher rubric levels are associated with higher levels of achievement in pilot elementary schools.

Middle Schools

For the secondary school analyses, one of a student's teachers in a subject related to the assessment measure was chosen at random. Since a student could have up to ten teachers over the course of the school year, it is likely that many of the students had teachers with different rubric levels. This biases the analysis against detecting a statistically significant difference between rubric levels. Despite this bias, the middle school pilots exhibit a positive relationship (i.e., achievement rises as rubric level rises) between rubric level and achievement in four out of six tests.

- At the middle school level, achievement increases with rubric level on the *ITBS* Math and the *CSAP* Reading and Math tests, with statistically significant differences on the *ITBS* Math and *CSAP* Math test.
- For *ITBS* Reading the relationship is reversed, with level 2 significantly higher than level 3 and higher than level 4.
- For *CSAP* Writing, students of rubric level 3 teachers outperform both level 4 and level 2, but the differences between levels are not statistically significant.

High Schools

At the high school level, the positive relationship between higher average NCE scores and higher rubric levels is also evident:

- At Manual High School on three of the tests (*ITBS* Reading, *CSAP* Reading and *CSAP* Writing), achievement increases with the level of rubric, with significant results on two of the tests. At Thomas Jefferson High School, the positive relationship is found on the three *CSAP* tests, with rubric level 4 on the Writing test significantly higher than level 3.
- At Manual High School there was almost no difference between levels on the *ITBS* Math exam.
- On the *ITBS* Language and *CSAP* Math tests at Manual and on the *ITBS* Reading and *CSAP* Reading tests at Thomas Jefferson the relationship is mixed, with students of level 3 teachers performing higher than students of level 4 and level 2 teachers. Significant differences are noted on only one of these tests.

Like the middle school analysis, the high school analysis is biased toward finding no rela-

tionship. That we find a statistically significant positive relationship between rubric level and student achievement on three of the subtests at the pilot high schools, and that most of the non significant results show the same relationship, is evidence that the relationship holds at the high school level as well.

Impact of Teacher Characteristics on Meeting Objectives

By their own measures and bodies of evidence and as verified by the building principals, teachers reached their objectives at a very high rate. The "met" or "not met" data collected by the district over the four years of the pilot show that 89% to 93% of the teachers met one or more objectives and were compensated. *Figure 4-10* shows the numbers and percentages of objectives that were met or not met by rubric categories.

Turning to the subset of objectives written by classroom teachers, one sees in *Figure 4-11* that a classroom teacher's educational background and experience are related to whether classroom teachers accomplished their objectives. TIRs met

FIG. 4-9

Estimated Mean NCE by Rubric Level, Adjusting for School and Student Characteristics (at the High Schools Student Characteristics only) Estimated from HLM and LSR Models Presented in Chapter VI

	Rubric Level	ITBS Reading	ITBS Language	ITBS Math	CSAP Reading	CSAP Writing	CSAP Math
Elementary Schools	4	50.9*[1]	56.8*[1,2,3]	39.7*[2,3]	54.6	52.4	56.5*[2,3]
	3	49.7*[1]	44.6*[4]	47.1*[1,4]	54.8	52.0	52.9*[4]
	2	49.5*[1]	43.6*[4]	47.0*[1,4]	54.2	51.9	52.5*[4]
	1	42.7*[2,3,4]	39.9*[4]	36.9*[2,3]			
Observations		8554	5324	6825	4556	5597	2127
Middle Schools	4	33.3	42.2	38.3*[2]	44.0	42.8	53.1*[2,3]
	3	33.4*[2]	41.6	35.9	43.5	44.4	47.6*[4]
	2	34.9*[3]	42.1	34.8*[4]	43.6	43.7	48.1*[4]
	1						
Observations		1789	1433	989	2238	2263	1693
Manual High School	4	40.2*[2]	34.2	37.2	43.2	38.6*[2,3]	33.8*[3]
	3	37.6	36.0	37.5	42.1	33.9*[4]	37.7*[4]
	2	37.0*[4]	34.2	37.7	41.8	35.8*[4]	35.9
	1						
Observations		675	415	556	685	331	491
Thomas Jefferson High School	4	55.7		54.3	57.8	58.2*[3]	60.8
	3	57.4		55.7	56.9	55.5*[4]	57.3
	2	55.7		56.7	57.2	55.8	59.4
	1						
Observations		1136	No Testing	807	920	471	706

*[1] = different from Rubric 1 at p< 0.05
*[2] = different from Rubric 2 at p< 0.05
*[3] = different from Rubric 3 at p< 0.05
*[4] = different from Rubric 4 at p< 0.05

FIG. 4-10

All Objectives—Rubric Levels by Met/Not Met Status by Year, 1999-2003

Year	Status	Rubric Level					Percent of Total			
		1	2	3	4	Total	1	2	3	4
1999-2000	Met	50	392	157	6	605	8.3	64.8	26.0	1.0
	Not Met	2	26	7		35	5.7	74.3	20.0	
	Total	52	418	164	6	640	8.1	65.3	25.6	0.9
2000-2001	Met	88	397	158	64	707	12.4	56.2	22.3	9.1
	Not Met	12	24	14	6	56	21.4	42.9	25.0	10.7
	Total	100	421	172	70	763	13.1	55.2	22.5	9.2
2001-2002	Met	7	570	379	156	1112	0.6	51.3	34.1	14.0
	Not Met	4	84	51	12	151	2.6	55.6	33.8	7.9
	Total	11	654	430	168	1263	0.9	51.8	34.0	13.3
2002-2003	Met	4	313	518	322	1157	0.3	27.1	44.8	27.8
	Not Met		27	39	31	97		27.8	40.2	32.0
	Total	4	340	557	353	1254	0.3	27.1	44.4	28.1

83% of their objectives while regular classroom teachers met 92% of their objectives, and first year teachers were less likely to meet their objectives than more experienced teachers. As mentioned earlier, inexperienced teachers could use assistance in the objective setting process; here one sees that providing support in meeting the objective would also be of value.

Teachers with tenure of 15 or more years in the DPS system are less likely to meet their objectives than teachers with under four years, four to 10 years, or 11 to 14 years of experience. This finding is consistent with other research that indicates that while inexperienced teachers (under three years) are typically less effective than more experienced teachers, the benefits of experience eventually begin to level off and may begin to decline before the twentieth year.

As teachers gain more years of experience in the pilot, their chances of meeting their objectives increase significantly. Eighty-nine percent of the objectives of first year pilot participants were met, by the fourth year of participation the success rate rose to 98%.

Student Achievement and Teacher Success in Meeting Objectives

This analysis revealed evidence of a positive relationship between the total number of objectives (out of two) a classroom teacher met and student achievement. Mean achievement scores were estimated for elementary and middle school students by the number of objectives met by their teachers, adjusting for student and school characteristics. In addition, mean scores by number of objectives met were estimated separately for each pilot high school, adjusting for student characteristics. This analysis is discussed in full in Chapter VI and is summarized in *Figure 4-12*. At Manual and Thomas Jefferson High Schools, there were no students with both achievement scores and a teacher who met either one or no objectives in some years for some tests. Indeed, for all tests the Thomas Jefferson High School analysis compares meeting one objective to meeting two objectives.

Elementary Schools

At the elementary school level, students of teachers who met both objectives had higher average

FIG. 4-11

Classroom Teacher Objectives—Percent Met By Teacher Characteristics, 1999-2003

Teacher Characteristic	Met Objective Percent (n)	Did Not Meet Objective Percent (n)
Teacher-in-Residence		
No	92% (1438)	8% (129)
Yes	83% (109)	17% (22)
χ^2 = 10.9, p=0.001		
Educational Degree		
Bachelor's	88% (796)	12% (109)
Master's	95% (663)	5% (34)
Doctorate	83% (5)	17% (1)
χ^2 = 25.2, p=0.001		
Years of Experience in DPS		
0 to 3	95% (350)	5% (19)
4 to 10	90% (221)	10% (24)
11 to 14	95% (394)	5% (19)
15 or more	85% (403)	15% (69)
χ^2 = 35.6, p=0.001		
First Year Teachers		
First Year	86% (73)	14% (12)
Subsequent Years	92% (1323)	8% (114)
χ^2 = 4.0, p=0.044		
Years of Pilot Participation		
1	89% (542)	11% (65)
2	93% (385)	7% (29)
3	94% (228)	6% (14)
4	98% (121)	2% (3)
χ^2 = 13.5, p=0.004		

scores then students of teachers who met only one objective:

- The difference in mean scores was statistically significant for all of the tests except *CSAP* Writing.
- On three of these tests, *ITBS* Language and *CSAP* Reading and Math, the achievement scores of students whose teachers met two objectives was also statistically higher than students whose teachers met no objectives.
- For the remaining two tests (*ITBS* Reading and Math), the mean achievement scores were not statistically different whether the number of objectives met was two or zero.

There is clearly an association between higher average NCE scores and meeting two objectives

compared to meeting one objective. However, the relationship may be more complicated when comparing meeting two objectives to meeting no objectives. Very few teachers met no objectives, making it more difficult to detect a significant difference between meeting one or two objectives and meeting no objectives.

Middle Schools

The middle school results, as expected, are less definitive than the elementary findings:

- On the *ITBS* Language and Math and the *CSAP* Math tests, having met one or two objectives produced similar mean student achievement, while meeting no objectives was associated with lower student achievement. The difference between meeting one or more objectives and meeting no objectives was statistically significant only for the *ITBS* Language test.
- On the *ITBS* Reading test students of teachers who met one objective had significantly lower mean scores than students whose teacher met either two or no objectives.
- For *CSAP* Reading and Writing there is no difference in achievement associated with number of objectives met.

Despite the bias of the statistical methodology against finding a relationship at the middle school level, there is evidence that students of teachers who met one or two objectives had higher average student achievement than students of teachers who did not meet any objectives.

High Schools

Turning to the high schools, similar results to the middle schools can be seen:

- At Manual High School on *ITBS* Reading, students of teachers who met two objectives have significantly higher scores than students of teachers who met one or no objectives. On the *ITBS* Math, and *CSAP* Writing tests the average score of students whose teacher met two objectives is higher (but not statistically significant) than the average score of students whose teachers met one objective. In addition, although the difference is not statistically significant, on the *CSAP* Math test average scores for students of teachers who met two objectives are higher than for students of teachers who met no objectives.
- At Thomas Jefferson High School, on the *ITBS* Reading exam, students of teachers who met two objectives had a significantly higher average score than students of teachers who met one objective. On the *ITBS* Math and *CSAP* Reading and Math tests, the students of teachers who met two objectives had higher average scores than students of teachers who met one objective, however the differences were not statistically significant.

Statistically significant differences are seen only for the *ITBS* Reading test (at both pilot high schools) between the average scores of students whose teachers met two objectives and the average scores of students of teachers who met one objective. However, the fact that the findings are biased against finding a statistically significant result, coupled with the fact that six of the remaining high school tests exhibit a positive relationship between number of objectives met and average achievement scores, make it reasonable to conclude that at the high school level meeting two objectives is associated with higher average student achievement levels than meeting one objective.

With a small number of exceptions, at the elementary, middle, and high school levels, higher average student achievement is associated with teachers who met two objectives compared to students of teachers who met one or no objectives. Of the 22 subtests examined (six elementary, six middle school, and 10 high school), four tests showed that students of teachers who met two objectives had significantly higher mean scores than students of teachers who met either one or no objectives, three tests showed a significantly higher mean comparing meeting two objectives to meeting one objective, and one test showed a significantly higher mean comparing meeting two objectives to meeting no objectives. Due probably to the small number of observations in the met zero objectives category, three of the subtests just mentioned show that students of teachers who met no objectives had approximately the same average scores as students of teachers who

met two objectives, while students of teachers who met one objective had significantly lower average scores. In addition, of the ten tests which did not exhibit statistically significant differences, nine more tests appear to show a relationship between higher mean scores and meeting two objectives while only one appears to show a negative relationship.

Setting objectives that garner a four on the rubric is not likely in and of itself to produce more learning. However, setting an excellent objective as the first step in a reflecting, planning, teaching, assessing loop that is carried out recursively and meta-cognitively by the teacher is a more persuasive explanation. The association between higher quality objectives and higher average student achievement on independent assessments, along with the positive association between a teacher's meeting two objectives and higher average student achievement, provide two of the most promising findings of the study.

Pilot and Control School Improvement Plans and Control School Teacher Goals

In an effort to further understand the institutional influences on teacher-written objectives, the study examined school improvement plans in both the pilot and control schools and the control school teacher goals. School plans provide insight into the general focus of a school and the areas identified for improvement by school councils. The control school teacher goals provided a picture of the process used throughout the district prior to the

FIG. 4-12

Estimated Mean NCE by Number of Objectives Met, Adjusting for School and Student Characteristics (at the High Schools Student Characteristics only) Estimated from HLM and LSR Models Presented in Chapter VI

	Objectives Met	ITBS Reading	ITBS Language	ITBS Math	CSAP Reading	CSAP Writing	CSAP Math
Elementary Schools	2	49.5*[1]	45.5*[0,1]	47.0*[1]	54.8*[0,1]	52.1	54.0*[0,1]
	1	47.4*[2]	43.6*[2]	43.7*[0,2]	52.7*[2]	51.5	50.1*[2]
	0	48.1	43.2*[2]	47.0*[2]	52.6*[2]	52.0	45.5*[2]
Observations		8608	5412	6870	4556	5609	2117
Middle Schools	2	33.9*[1]	40.7*[0]	35.0	43.4	45.1	46.8
	1	32.1*[0,2]	41.4*[0]	35.0	43.3	45.6	46.5
	0	35.0*[1]	37.6*[1,2]	33.7	43.2	45.1	44.1
Observations		1800	1453	1011	2223	1325	950
Manual High School	2	37.0*[0,1]	32.2	38.0	42.5	36.8	37.4
	1	33.3*[2]	37.0		42.9	34.7	
	0	33.1*[2]		35.9	42.7		35.6
Observations		689	428	585	687	333	510
Thomas Jefferson High School	2	57.1*[1]		55.7	57.0		60.2
	1	51.7*[2]		54.1	55.6		58.3
	0						
Observations		1137	No testing	809	917		704

*[2] = different from Met 2 Objectives at p< 0.05
*[1] = different from Met 1 Objectives at p< 0.05
*[0] = different from Met 0 Objectives at p< 0.05

implementation of the PFP process in the pilot schools. These artifacts were examined and compared at two points in the pilot: Spring 2001 and Spring 2003. The analysis of school plans is discussed in the context of organizational alignment in Chapter VIII of this report.

There is a summary of the findings from the review of 12 sets of control school teacher goals for the school year 2001-2002 in the mid-point report (pp. 36-37). The control teachers wrote three annual goals, two of which were academic and focused on school and district goals. The third goal was optional, but most teachers used it to set a professional or personal goal. Teachers wrote goals on the form provided and then weighted them, giving priority to some goals over others or treating them equally. Other sections of the form asked for teaching strategies and provided an appraisal section for the principal. Most of the objectives were written in an assessment-focused manner with a percentage of attainment; however, there was almost no use of baseline data.

The reading of 19 sets of control school goals in Summer 2003 showed some changes in the formats that teachers were using: 12 schools used the standard district form that was observed in 2001; six schools used the PFP process and forms or modified versions (e. g., the PFP template worksheet with the objectives transcribed to the standard district form); and one school provided a two-page summary listing one or two goals per teacher without teaching strategies. Some of the standard forms had removed the appraisal categories. Except for those using the PFP process, there was generally not a reference to baseline data or starting points for students.

The control school goals reflect the objective-setting process that pilot school teachers had used prior to the inception of the pilot and help explain the differences and difficulties of setting PFP objectives experienced by pilot teachers. Also, as noted, in the mid-point report, they helped explain the preference for assessment-focused objectives in the initial years of the pilot.

The control school goals reviewed in 2003 show the migration of features of the pilot, in this case, the objective format, into the control schools. This circumstance is not unexpected in light of interviews where several control school principals indicated admiration of the pilot process, particularly the objective setting component. Secondly, schools other than pilot schools were introduced the use of the OASIS system for accessing student data. Based on this sample, however, little has changed in control school goals since 2001, except in the 25% of this sample that have begun to use the PFP processes.

E. Summary

The design of Pay for Performance in Denver is centered on the outcome of two teacher-developed objectives. When the objectives are met, additional compensation is earned. Because of this pivotal role in the design of the pilot, objectives became a key element not only of the implementation but also of the study of PFP. This chapter explained that the objectives for the pilot were grounded in past practice in the district, but new features and expectations for the objectives made a familiar way of doing things more complex—creating conflict between the old and the new in the objective setting process.

A four-level rubric was developed to measure the quality of the objectives and a set of questions that explored the relationship of the objectives to several other sets of data was also developed. The analyses of the objectives over the four years show that (1) learning to write objectives for the purpose of establishing and achieving growth targets for students is more complicated for teachers than might have been expected; (2) the setting of objectives, nonetheless, improved over the life of the pilot as technical assistance improved and experience increased; (3) inexperienced teachers (first year and TIRs) would benefit from additional assistance in developing and implementing their objectives; (4) higher rubric levels are associated with higher average NCE scores on independent measures; and (5) meeting two objectives is associated with higher average student achievement than meeting only one objective.

These findings point to the impact that objective setting has on student achievement and show the potential for objectives to be the basis of a component in a compensation system and the springboard to improved student achievement. Setting objectives that garner a four on the rubric

is not likely, by itself, to produce more learning. However, setting an excellent objective as the first step in a loop of thinking, planning, teaching, and assessing that is carried out recursively by the teacher is a more persuasive explanation. The positive relationship between higher quality objectives and student achievement in most areas of independent measures, along with the positive relationship between number of objectives met and higher average student achievement, provide two of the most promising findings of the study. These findings suggest an agenda for professional development not only in Denver but also in other districts initiating achievement reforms. In the upcoming chapter, the role and impact of objectives are explored more thoroughly through the perspective of pilot teachers.

CHAPTER V

The Teacher Perspective

A. Introduction

Launching the Pay for Performance pilot from an established district practice of goal-setting promoted participation and a quick start-up. Doing so also added to the complexity of implementation and affected how pilot teachers perceived their roles and obligations. Although building on a pre-existing practice had a double-edged influence on the pilot, the data indicate that this objective setting process holds promise for the district. There are statistically significant positive correlations between teacher objectives rated at higher levels on the research rubric and student achievement on most sections of the independent assessments. Further, there is a positive correlation between teachers' meeting two of their objectives and student achievement on these same measures. These results make understanding and responding to pilot teacher perceptions about objectives not only worthwhile but also essential for any future work connecting teacher and student performance for compensation purposes.

The evolving quality and impact of the objectives, as well as teacher descriptions of their work with objectives, provide a barometer of the teacher experience of the pilot. Besides setting objectives and assessing yearly progress on objectives for purposes of additional compensation, teachers have participated in the study of the pilot by responding to annual surveys and interviews. Additionally, smaller samples of pilot teachers participated in focused interviews and focus groups. Several opened up their classrooms for detailed observations. Over the four years of the study, pilot teacher input has been voluminous and has contributed significantly to the key findings of the pilot.

In the spring of each year of Pay for Performance (2000-2003), teachers in the pilot schools, as well as other stakeholders in the pilot, responded to surveys. Additionally, a random sample of the groups was interviewed each spring. The

protocols for both surveys and interviews for pilot teachers focused on the impact of various elements of the pilot on teaching and learning, as well as other changes during the respective year that might be attributed to the pilot. As the pilot progressed, survey questions were designed to validate perspectives that emerged in the interview data. Themes about the impact of the pilot on student achievement emerged from these data. Most of the themes that were apparent after two years and that were reported at the mid-point of the pilot, remained at the end of four years but were more thoroughly articulated and better understood. Teacher responses to the challenges of the objective setting process have been more descriptive, analytical, and solution-oriented in the latter years of the pilot than in the beginning years.

Teaching in Denver, a large urban school district with a diverse student body, has inherent challenges. During the span of the Pay for Performance pilot, teachers were also becoming acquainted with the new state assessment, the *Colorado Student Assessment Program (CSAP),* which was implemented in grade level segments over time and which became a public report card on the performance of their respective schools. Other large scale reforms and programs that affected teachers included the implementation of an area organizational structure, the development of three small schools from one of the pilot high schools, and a new literacy program mandate for most of the schools in the final year of the pilot. Simultaneously, as described in Chapter VII, Denver experienced personnel changes at the district and pilot school leadership levels, and several hundred new teachers were inducted each year into the entire district. Teachers had input into few of these changes. They were, though, able to choose to become part of the pilot. Their decision to do so started the teachers down a new and largely uncharted path. Certainly, Pay for Performance did not come with a roadmap that would lead teachers to expect to make fundamental changes in their practices in order to impact student achievement.

Therein lies one of the central stories of Pay for Performance: how teachers understood and responded to the goals of the pilot, which had a simple design for reaching a complex outcome—improving achievement student by student.

As both the agents of the pilot and key subjects of the study, teachers in the pilot schools have held a unique position as knowledgeable critics of the process. They have provided four years of critique and feedback through interviews and surveys. Additionally, during the third and fourth years of the study, representative groups of teachers engaged in special in-depth components of the research. Vehicles for doing so included (1) a set of focused interviews wherein 12 teachers describe their objective setting process from beginning to end, and (2) a deeper qualitative study of how 16 teachers in four pilot schools experienced the pilot. Through these efforts, teachers provided their insights into the process and their interpretations of some of the survey and interview findings that, on the surface, seemed contradictory. Thus, teachers have been de facto researchers as well. This chapter pursues a deeper understanding of the teacher experience of the pilot through surveys, interviews, and two qualitative studies that included focused interviews, group interviews, and classroom observations. Finally, this chapter examines the ideas and suggestions tendered by teachers for improving the processes of Pay for Performance.

B. The Intent, Impact, and Process of Setting Objectives

Teacher Understanding and Descriptions of the Intent and Impact of PFP

Clearly, one intent of Pay for Performance was to increase student achievement by providing additional compensation to teachers for meeting student growth targets. However, there is not necessarily a direct link between setting an objective with a growth target and increased student achievement. Further, the design of the pilot did not provide a blueprint for what should happen between setting objectives, based on student achievement data, and collecting evidence of meeting those objectives. It is a design respectful of teacher autonomy. It is based on the assumption that, in setting more informed objectives and being accountable for the outcome, teachers will make any necessary changes in classroom practice.

Working with a familiar process and maintaining a degree of autonomy allowed pilot teachers to engage in what could have been a high stakes

program with a minimum of risk or commitment to fundamental change. Based on what teachers have said over the course of the pilot about their reasons for joining the pilot and the actual impact of PFP on their teaching, most did not plan to change what they were doing in the classroom when they first entered the pilot.

According to Spring 2000 survey data, some teachers intended to get the bonus. As one teacher comments, "We joined the pilot to get a bonus for what we already do." Another adds, "We have been setting goals for many years. We are a school with excellent teachers. Why not get paid for the extra hours we work?" A third teacher indicates, "I felt I would at least get $500 for trying. I did not change my teaching. I feel I work hard whether there [is] PFP or not."

Others thought of joining the pilot as doing one's professional duty. One teacher suggests, "Being part of the pilot would help me to better evaluate and have input into the program before it is ruled on by teachers." Another notes, "I wanted to have the experience of PFP so I would have valid information when voting."

Other less prevalent reasons for joining the pilot included the influence of the building principal, the opportunity for professional development, the satisfaction of intellectual curiosity, and lastly, a belief that PFP might improve student achievement.

Some teachers reported that they were told either by their principals or by some representatives of the Design Team that they did not have to make changes. Since most teachers believed that they were already giving their best, such statements gained currency among many teachers. Nonetheless, survey and interview data show that most pilot teachers *did understand the goals of PFP* at the outset. In the Spring 2000 survey, not long after the pilot was underway, 85% of pilot teachers agreed or strongly agreed with the statement that a goal of PFP is to "increase student achievement." Seventy-five percent agreed or strongly agreed that another goal of PFP was to "focus district activity on improving teaching and learning," and 78% agreed or strongly agreed that increasing "teacher accountability for student achievement" was a goal.

Early on, teachers *did* realize that setting objectives for compensation was not exactly business as

FIG. 5-1

Project Support Needed for Pilot to be Successful, 2000

Project Support	Strongly Agree/ Agree	Strongly Disagree/ Disagree	N	Rank*
Training in objective setting	68.4%	31.6%	345	9
Greater access to student achievement data	70.9%	29.1%	344	8
Better understanding of student achievement data	71.4%	28.6%	343	7
Help in developing and implementing new teaching strategies	70.9%	29.1%	340	6
Feedback on the success of my methods	84.8%	15.2%	341	1
Greater clarity on how objectives should be set and measured	83.3%	16.7%	342	2
Ways to set objectives based on the needs of my students	81.1%	18.9%	339	4
More time to analyze data and develop my skills	82.2%	17.8%	342	3
Greater access to technology to analyze student achievement	72.1%	27.9%	341	5

*Based on percent strongly agree/agree

usual. They quickly identified areas where greater knowledge, access, and support would improve their objective writing. There was training in setting objectives that particularly focused on these new elements, but as some teachers would later report, "Not enough." The Spring 2000 survey responses show that teachers believed that, for the pilot to be successful, participants would need: greater clarity on how objectives should be set and measured (83%); greater access to technology to analyze student achievement (72%); and more time to analyze data and develop their skills (82%). Additionally, 81% indicated that they needed ways to set objectives based on the student needs. (See *Figure 5-1*) Thus, there was early recognition that setting objectives for compensation purposes did require changes from the prior district practice, most notably, the use of baseline data and the projection of a growth target for a class of students. Further, there would be consequences for the teacher's performance in relation to the objectives. Teachers asked for more assistance to build their capacities in these areas.

Near the end of the pilot (Spring 2003), many teachers identified positive impacts (See *Figure 5-2*) of the pilot in the areas of setting expectations for students, having access to data, and understanding and using student data in setting objectives and planning. These are the areas that they had identified in 2000 as ones where they needed more information and support. Thus, from their own descriptions, most teachers understood the goals of PFP to be the improvement of teaching and learning. They came to recognize the importance of using achievement data to measure student growth, wanted more assistance in working with data and establishing student expectations, and ultimately believed that they or their schools had experienced positive impacts from participating in the pilot. These teacher perceptions are discussed in greater detail in Chapter VII. Few identified negative impacts and approximately one-third identified no impact. The item with the lowest positive impact—the availability of appropriate assessments—is an organizational issue that has plagued the pilot and is addressed in Chapter VIII.

Perceptions about the Impact on Student Achievement Areas

Pay for Performance, as noted above, had generally well-understood goals for improving student achievement, as well as a required format and process for objective setting that prompted teachers to use student achievement data more effectively to identify the baseline and measure growth. A charge of the pilot study was to follow the impact of the pilot on student achievement, not only through analysis of the achievement data but also through teacher perceptions. From the feedback provided by pilot teachers in the first two years of the pilot in interviews, the theme of "greater focus on student achievement" emerged as one major response to questions about the impact of

FIG. 5-2

Identification of Impact of PFP, Spring, 2003

Areas of Impact	Positive Impact	Negative Impact	No Impact	N
Expectations that I set for my students	63.5%	0.6%	35.9%	345
Timely access to student achievement data	62.2%	4.4%	33.3%	339
Understanding of student achievement data	64.9%	0.9%	34.2%	339
Use of student achievement data to set objectives	66.1%	0.6%	33.3%	342
Use of student achievement data to plan instruction	59.9%	0.6%	39.5%	337
Availability of appropriate assessments to measure growth of my students	54.1%	2.7%	43.2%	338

Pay for Performance. In subsequent years, the focus on student achievement for the teacher and the school strengthened; it became the most frequent response to questions of impact.

The following sample of teacher responses from Spring 2003 interviews demonstrates that pilot teachers became more articulate about what "greater focus" meant to them personally and in their schools, particularly the impact on school culture:

> "With PFP, you don't forget the goals, and it is possible to be more consistent over the course of the year."–Pilot teacher

> "I'm placing my ideas on paper and that is very helpful to me. By being more formal and deliberate about my objectives, I find that I reflect more on the substance of the documents and it helps me communicate my efforts more clearly to others. PFP helps the staff focus on analysis rather than just assuming. It allows us to share more with each other at a deeper level."–Pilot teacher

> "I have been observing the dialogue around PFP and I think we have a better environment for students because of PFP in this school year. We started this year by looking at more data about our students and I think many teachers are checking the growth of their students more often. I think teachers are reflecting more upon the needs of their students than in the past."–Pilot teacher

> "There is a lot more teamwork. We are on the same page. There is a lot of discussion about school-wide goals. There is lots of focus on students who are borderline or below grade level."–Pilot teacher

Survey data also show that there is a significant difference between the percentage of respondents in Spring 2002, who felt that the *focus on student achievement* had stayed the same (31%) and those who felt that there had been an improvement or increase in focus (64%). This difference increased again in Spring 2003, with 68% reporting a positive impact or change in "my school's focus on student achievement." Twenty-nine percent reported that there had been no impact on the school focus. These perceptions are described further in Chapter VII.

Another area where respondents increasingly attributed a positive impact to Pay for Performance at a significantly higher level than to "negative impacts" or "no impact" is in *the expectations that I set for my students.* In the Spring 2003 survey responses, 64% identified a positive impact as compared to those who saw a negative impact (less than 1%) or no impact (36%). Thus, by the end of the pilot, the surveys identified three positive impacts in the student achievement area, each emerging from the data at a consistent and significant level:

- Increased focus on student achievement;
- The expectations set for my students; and
- A cluster of positive impact responses around the availability and use of student achievement data to set objectives and plan instruction.

These are the kinds of changes that should predict positive impacts on teaching and learning and lead to changes in teaching practice. Yet, one-half to two-thirds of teachers surveyed and interviewed over the life of the pilot have maintained that they have not changed their teaching. This apparent disconnect among the responses constituted one of the more puzzling aspects of the study and required additional probing to understand.

The ambivalence that pilot school teachers felt about the impact of PFP on their teaching is shown with 53% responding that PFP has had no impact on "my knowledge and skill in delivering instruction" in Spring 2003, though 63% had reported improvement in Spring 2002. Fewer than half of the pilot teachers surveyed saw a positive impact of PFP on their knowledge and skill in delivering instruction (47%) or their knowledge of subject matter (44%). As teacher research indicates, these are areas likely to impact student achievement.[1]

Teacher Descriptions of the Objective Setting Process

It was important to gain more insight from teachers themselves about how they engaged in the process of setting objectives and how that process affected their teaching. An interview protocol was designed during the Spring 2002 interview season for the purpose of asking a sample of teachers to describe their processes from beginning to end. Twelve

teachers were selected from the 64 randomly selected pilot teachers already scheduled for spring interviews. The 12 teachers did not know ahead of time that they would be asked to describe their processes with objectives. Of the 12 teachers, two had been in the district for two years, but had taught elsewhere, and another was a new teacher in his first year working on alternative certification. Nine of the 12 had been in the pilot schools for all three years of the pilot at the time of the interviews.

The interview protocol, administered by three different interviewers, was comprised of the five questions listed in *Figure 5-3* with prompts as needed:

The analysis of the objective-focused interviews was based on how teachers described key decisions during the process, what they learned during the process, what obstacles they confronted, and how they thought through the process and its potential outcome. Based on the similarities of the responses of teachers, three patterns of thinking about objectives were identified—*innate, purposive,* and *accountable.* The three groups of teachers, one for each pattern detected, are described and discussed below. There were five teachers each in *Groups One* and *Two* and two teachers in *Group Three.*

The five teachers in *Group One* are diverse: a first-year teacher, a third-year teacher, a mid-career teacher, and two teachers who have taught 25 or more years. They teach in four different pilot schools. The new teacher admitted that he made a "muddle of objective setting" and that "it was a low priority." All of the other teachers in this group can describe a full process, though their descriptions, until prompted for greater detail, are general or cursory in nature.

FIG. 5-3

Objective-Focused Interview Protocol

How do you develop objectives for PFP?
Regarding the objectives you set, what interactions do you and the principal have over the course of the year?
Has setting objectives under PFP had an impact on learning?
What kinds of support or professional development have been helpful or would be helpful in achieving the objectives that you set?
If you could change anything about how the PFP objective setting process is designed what would it be?

The major learning from the objective setting process for this group has been around the use of student achievement data and thinking about "reasonable growth," although they did not reference it as such. The objective process obstacle most mentioned was the determination of what constitutes reasonable growth and how to set a growth target that can be reached. Some descriptions about how they addressed these issues include:

> "I followed the Design Team recommendation for the first year (3% overall) and then adjusted it in subsequent years based on additional data."

> "I thought about what I could do if I worked really hard with the kids but without setting my goal so high that I have no chance of getting there."

> "[Setting growth targets is] pretty tricky because the levels on the *DRA* are not even." This teacher resolved this dilemma by taking a "cut-off" point of five levels beyond where each student started, but he tries to "take students as far as they can go in reading and math, so setting lower than 100% is not reflective of what I do."

The descriptions of their processes suggest that they relied on their teaching instincts to arrive at a reasonable growth target. Several implied that they were allowed to be "less stringent" on PFP goals so that they could meet them, and at least two mentioned this fact as a problem with the concept of the pilot. There is a tendency to blame the pilot for the setting of lower growth targets than they might have done without the pilot. One teacher expressed concern that setting only two objectives will "narrow the curriculum."

Except for the first year teacher, teachers in this group were certain that they had not changed their teaching as a result of PFP. One teacher explained that there has been "no altering of the curriculum." However, while he had thought writing goals a waste of time in the past, he is now "paying more attention because of the fact that I made a prediction about growth" and wants to see how it comes out. Another teacher was

"almost insulted" by the idea that she should change her teaching for PFP, a concept with which she disagrees. Still another teacher expressed "shock" that the pilot as she has experienced it is "not a negative, even though it is not a positive." Still another observation: "I think the [objectives] have had an impact on learning. I can't say that PFP has had an impact."

Most, including the very newest, struggle with difficult-to-teach students and seem to try to distance themselves from the student learning issues. One teacher, in describing options available for students who are not learning, noted that since parents are informed "there shouldn't be anybody who is surprised if they don't meet the mark."

In general, teachers in *Group One* believe in their own *innate* teaching abilities and are generally unscientific about objective setting. One explains that he has been "teaching reading successfully for a long time." The new teacher believes that if he can just spend more time with his students, they will learn (provided some other home conditions are met). Two of the teachers in this group make several "we" statements suggesting that some of their process may be collegial, but three do not. Goal or objective writing seems to be what one does to meet a requirement not what one does to focus one's teaching, engage with one's colleagues, or improve one's own performance.

Yet, all of the teachers in this group recognized that their use of data in setting objectives had improved. Using data to determine the baseline and measure progress is the part of objective setting about which this group of teachers said the most, but they did not connect this part of the process with student outcomes. For one teacher, following data on his students is like an athlete measuring his personal best. Still another teacher states the mixed response that is evident in so much of the perceptual data:

> "Has PFP had an impact on learning? Not really. Has it made me a better teacher? Not really. But it did help me to use data. We have always been doing the same thing. Now it's just on a piece of paper."

Each teacher in *Group One* pondered and resolved the issues of reasonable growth in ways that satisfied them, but none connected the exploration of student data—particularly schoolwide—with the potential for clearer or more scientific answers to the question of reasonable growth. Overall, *Group One* contained both teachers who are inexperienced and teachers who are well meaning but unable or unwilling to use reforms to think systematically about teaching.

The five teachers in *Group Two* are experienced, most having taught from 12 to 27 years, though one had only four years of experience. They teach in five different pilot schools. Two of the teachers talk in first person singular, while three speak as "we," describing school/grade level processes of reviewing data and considering growth and setting a focus. The interviews show positive talk about participation in the pilot and its potential for their students:

> "We embraced PFP from the beginning."
>
> "There is no doubt about our purpose and focus."
>
> "With all of this focus, the students have to be the beneficiaries."
>
> "I'm looking for more techniques that motivate children."
>
> "Students benefit indirectly from teacher growth."

These interviews contain evidence of more positive attitudes about learning from Pay for Performance and about interactions with other staff members, students, and the principal: "This staff is stable and loaded with master teachers." They are generally thoughtful about the objective process. A specialist notes that other teachers, in writing Pay for Performance objectives, must do "something similar to [what I do to write] my IEP [Individual Education Plan]." There is a mention in one instance of the need for multiple measures and for vigilance on the part of principals to avoid dishonesty, indicating some reflection about how to improve the process.

Group Two can be best described as *purposive* about objective setting. There is a positive sense of motivation and mission in the responses of these five, both from past experiences and from PFP, that teachers with "plenty of information" can focus on areas for improvement and that students will be the "beneficiaries." Giving specific examples of how they use data and what the limitations

of certain data are, they are more scientific about their use of data than the first group, though not necessarily scientific about teaching choices. There is greater motivation than skill in this group.

The two teachers in *Group Three* had each taught over 12 years and were working in two different pilot schools. They spoke as "we" teaching at schools where staffs are working together on assessment, screening, and gathering information to establish objectives. One spoke of an analytical tool that the whole staff uses: "We use a matrix to determine where we are strongest, where we need to fine tune, and where we need the greatest emphasis." In both schools, there is a collegial relationship with the principal (one new principal and one established principal).

Without prompting, they elaborated on the relationship of objectives to standards, of research about how students learn, and of research on teaching (pedagogy) as they describe their processes. For these two, students are the basis of their thought processes: "Students benefit from good teaching;" "We don't wait until the test results to recognize students who may be in trouble."

They did not claim that Pay for Performance objectives had led to major changes in their teaching, but they did know how objective setting fit into their planning and teaching processes, as evidenced in these statements from the two interviews:

> "Our objectives have been dictated by a change in our students and their families."
>
> "We use test scores and areas for improvement in our school's plan to help us know where we need to concentrate."
>
> "I don't know if PFP objectives are that much different than any others for teachers who are expected and determined to move children from one point to another."
>
> "We use research and staff development to determine what is good for students and various ways to teach."

These teachers were working in higher-performing pilot schools. Their interview responses show awareness that they are fortunate to have interested parents, but there is also a concern that their schools are in a district with overall low performance. One teacher indicates that her school cannot establish the baseline by district performance because the "curriculum for DPS is too low." Another remarked: "We dummied down how we wrote objectives to somewhat fit the DT [Design Team] examples." They attribute these issues to the lack of district curriculum leadership and the quality of the sample objectives provided, not to the district's students.

The process described by these two teachers may best be described as confident and *accountable.* They hold high expectations for themselves, for their colleagues, and for their principals in terms of doing what has to be done in order to succeed with students. They had been using data and setting goals and objectives based on available student data prior to the pilot and distinguish between the processes used before and the ones used for Pay for Performance. Confidence is built not only through knowledge and skills but also through practice, reinforcement, reflection, revision, and of course, success with students. Circumstances in their schools support continual growth in confidence, yet confident and accountable teachers also contribute to the supportive circumstances in the school.

Findings: Objective-Focused Interviews

- *Teachers brought different styles of thinking and sets of experiences to the pilot, impacting how they responded to the key requirements of writing two objectives, selecting assessments, setting growth targets, and conferring with the principal.* Descriptions provided by the 12 teachers show how the implementation of the pilot varied not only by school but by teacher, particularly based on styles of thinking about or making decisions about teaching. The descriptions also indicate the need for differentiated supports for teachers
- *Teachers who showed traits of accountability identified how objective setting fit into their planning and impacted their teaching.* They also had a history of higher rubric levels. Though this examination contains too small a sample to generalize about the relationship between the rubric levels and the way interviewees thought through the objective setting process, there is

other teacher research to indicate the relationship between teacher planning and accountability and effectiveness.

- *The role of the building principal in the effective implementation of the objective process was apparent in the teacher descriptions.* Among the teachers in *Groups Two* and *Three,* there were signs of positive attitudes, expectations, and interactions with their building principals even when they were new in the principal role. Except for the new teacher who "adored" his principal, the teachers in *Group One* could not articulate a clear process of interaction with their principals on objectives—or where they did, the interactions were more perfunctory than student-based. *Group One* teachers did indicate that they would like more interim feedback from the principal to know how they were doing, and they had vague notions that their principals would help them if they asked, which pales in comparison to the principals of schools who were noticed as frequent classroom observers and participants in grade level discussions of objectives. Often, the teachers in *Group One* were looking to the Design Team liaison for leadership and feedback, hoping for more support from that quarter or again believing that it was there for the asking even though they had not asked to date.
- *The importance of teacher dialogue and/or collaboration on the individual teacher's perception of their own processes was evident in both Groups Two and Three.* Those who were talking as "we" articulated a more thoughtful process and were more confident about their decisions. It was evident that they had talked through or explained their decisions and rationales in other venues—either with colleagues or principals—and needed little prompting to describe how they had approached and thought about the PFP objectives.

This set of interviews provides insights into what teachers brought to the objective setting process in the form of intellectual processes and expectations for themselves and their students and what they had received at that point in the form of new or deeper understandings, professional dialogues, and principal support.

In conjunction with other data, the teacher responses make the point that a "grass roots" approach to reform, that is, one which leaves the *how* of implementation to individual teachers, will succeed or not based on the skill, commitment, and accountability of those teachers and the commitment and support of the building principal. The study in the next section pursues this idea with a more structured and representative sample of teachers.

C. Changing Classroom Practice

As has been noted, the intent of the Pay for Performance pilot was to link teacher compensation to increased student achievement. The assumption that underlies the design of the pilot is that as an outcome of setting an objective and potentially earning a bonus, *teachers will teach differently,* and concomitantly, student achievement will improve.

Yet when asked about the impact of PFP on teaching practices, pilot teachers often responded in surveys and interviews over a period of four years that they had not changed their teaching practices in order to attain their objectives. A typical comment is "I'm not doing anything differently." However, responses to other survey and interview questions indicate that, for a preponderance of teachers, there was an increased focus on student achievement, an increase in the understanding of student achievement data, a greater use of student achievement data in planning, and an increased understanding of the need for greater alignment between objectives, instruction and assessment. Many practitioners would say that these changes *are* doing things differently and that the nature of changes identified are ones recognized in research literature as potentially contributing to increased student achievement.

These outputs of teacher processes (objectives, changes in teaching practice, uses of assessments, attainment of results, student achievement) had been largely understood through three years of teacher interview and survey data as well as available artifacts and documents. To augment the understanding of teacher processes, the final year of the study added a special component in addition to collecting and analyzing the fourth year of

perceptual data. Notably, CTAC conducted a qualitative study with 16 teachers, seeking a deeper understanding about the relationship between Pay for Performance and changes in teaching practices. *Figure 5-4* demonstrates the basics of the PFP processes with the shaded box showing the area of interest for the qualitative studies:

In the fall and winter of 2002-2003, the deeper qualitative study was designed to gain more teacher perspective on how PFP impacts teaching. The focus was on learning more from 16 teachers in four representative elementary pilot schools about how Pay for Performance affects teaching and learning.

Sixteen Teacher Study

Ten pilot elementary schools were eligible for the deeper qualitative study based on the number of years in the pilot and the elementary grade span. The ten schools were ranked on the selection criteria and four schools were identified as the most representative of the range of pilot elementary schools.

FIG. 5-4

Pay for Performance Processes

Objectives
Each teacher writes two objectives and provides evidence of student attainment to the principal.

PFP Study
Each objective is evaluated along with other related documentary evidence, student achievement data, and interviews and surveys.

Teacher Practices and Student Achievement
How does writing objectives lead to changes in practice that may increase student achievement?

Principal Decision
Teacher either meets or does not meet objectives. (Around 90% currently meet objectives.)

Cost
Teacher receives extra pay.
Cost to district = $855,250 (01-02)

PFP Study
Does student achievement improve on independent measures? How does pilot improvement relate to objectives? to control school improvement?

School selection criteria for the study included: (1) math, reading, and writing performance on the *Colorado Student Assessment Program (CSAP);* (2) school demographics (English language proficiency, free/reduced lunch, mobility, ethnicity); and (3) teacher demographics (years experience, years in the school, ethnicity, mobility, advanced degrees). Four teachers from the four schools were invited to participate. The names of the schools and participants are not used in this report.

Three classroom teachers and one specialist/special subject teacher/special education teacher in each of four elementary schools selected as representative of the elementary pilot schools were chosen. The selection criteria included: (1) their years in the pilot (no fewer than two) and (2) their potential to add new voices (not teachers who had been interviewed in the last year). The four specialists/specials were selected based on full-time assignments in the school in special education classrooms or in subject matter classes, such as music and physical education. The selection also allowed for observing ability grouping (English learners and gifted), heterogeneous grouping, and pull out instruction. The group also included an alternatively certified teacher.

On three different visits, the research team (1) observed all sixteen classrooms or workspaces for one full day; (2) conducted four 90-minute after-school focus groups, comprised of the teacher participants; (3) made two additional partial day visits to the classes; and (4) conducted a second round of 90-minute focus groups, comprised of the same teacher participants. This study involved a total of 12 hours of focus group interactions and more than 160 hours of classroom observation.

Findings: Sixteen Teacher Study

The key findings of the deeper qualitative study—as related to a positive relationship between the pilot and changes in classroom teaching practices—indicate that:

- *The teachers in the study did not interpret the pilot implementation as requiring changes in their core teaching practices in order to improve student achievement.* They were, in fact, told by Association representatives and at least one principal that they could earn a bonus for "doing what

they already [did]." Another principal encouraged the staff to participate, reasoning that pay for performance could be the wave of the future; thus the school should get into the pilot and find out about it. Thus, as a reform, it was something to try out and find out about.

- *The teachers in the study implemented the mandated elements of PFP: write two objectives based on baseline data, set growth targets, assess, and provide evidence of attainment to the principal.* These elements are the fundamentals that all teachers had to complete in order for their schools to be a participant. At least half of the teachers in the study indicated that the objective setting process had been more onerous than expected, particularly in "paper work," that changes in the format for objectives had come about each year, and that the availability of assistance has been inconsistent. Several had problems with the technology when entering their objectives. Nonetheless, they all tried to comply and meet the requirements of the mandate. All 16 teachers agreed that setting objectives for student growth based on baseline data is what they should be doing, though several said their school would have used baseline data anyway or, in the case of one school, were already using baseline data without a bonus.
- *Most of the teachers in the study did think that they have had better access to data and that they were currently using student data more systematically as a result of PFP.* However, three teachers of the 16 were adamant that any changes in the way that they use student data are attributable only to the *Colorado Student Assessment Program,* which pressures them to improve student achievement scores. Others saw both PFP and *CSAP* as influential on their use of data. The OASIS system was valued but its development was not attributed to the pilot. Interestingly enough, the teachers in the group that seemed to engage with the student data most readily and to see potential for the impact of better student assessment data on their teaching were special subject teachers: (1) a special education teacher sees the PFP process in light of the individual education plan process and helps other teachers in the school write measurable objectives; (2) a physical education teacher measures students (large numbers of them) by his own written and performance assessments, but follows their reading and math scores to see if there are relationships; and (3) the GATE teachers philosophically prefer authentic assessments, but understand that doing well on standardized tests is important to their students and may be a quality indicator of their program.
- *The teachers in the study were articulate about why and when they do change their core teaching practices.* As teachers in the first focus group sessions talked about why and how they do or do not change their teaching practices, a model of concentric circles emerged: the core practices in the inner circle, primary or immediate influences on core practices in the second circle; and secondary or potential influences in the outer circle. In the second focus group session, participants reviewed the model, made revisions, and elaborated on it.

Levels of Influence on Core Practice

Figure 5-5 shows the graphic developed with the teachers in the focus group to help explain why and when they make changes in their practices in order to implement new programs or goals.

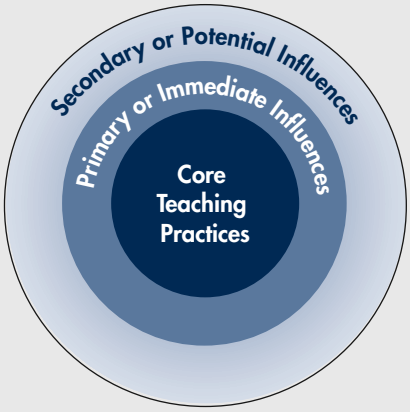
The graphic describes how teachers in the qualitative study talked about their practice and what has influenced them to change. Items in Level One are the most likely to penetrate core practice. The teachers in the study are relatively experienced, and as they suggested, new teachers might place some of the elements in different categories. For example, teacher evaluation may be more influential on an untenured teacher.

Pay for Performance and the Levels of Influence Schematic

Where does Pay for Performance reside in this schematic? Level One or Two? As teachers in the study describe the influences on their teaching practices, PFP resides in Level Two, not immediately influential on core teaching practice. How-

FIG. 5-5

Level of Influences on Change in Teaching Practice



Secondary or Potential Influences that may Change Teaching Practice
This level contains elements more removed from what teachers see as their core work. The elements may actually be completely external to a teacher's situation or they may be internal elements from which teachers distance themselves. If such elements are internal, they may be described as "administrative or paper work" or be perceived as something that has to be done but which has little or no impact on day-to-day teaching practice, such as teacher evaluation; standardized assessments (without accountability); objective setting; parent/school governance structures. General reforms that may be identified as good ideas but do not seem relevant to their situation or for which they do not have enough time in the day may also be in this area. They may also dismiss such reforms, believing or saying "I/we already do that."

Primary or Immediate Influences that are Changing Teaching Practice
This first level can be described as an internal change environment that contains elements pressing the teacher's practices such that there is a likely chance of penetrating the core. The teachers in the study identify influential elements in this environment as the following: agreements with colleagues (grade level articulation); new school focus, which may be articulated by the principal or come from the school plan or both; district-mandated programs; state-mandated assessments that label school performance; and new textbook series. Teachers may be trying out new practices, evaluating and incorporating them or parts of them into their core practice. These new practices may look different as integration occurs or they may become add-ons to the core practice.

Core Teaching Practice
The core teaching practices, described by the teachers in the study as "my style," "the way I do things," "what I know works," and "I know what's best for my kids," represent sets of teaching strategies (pedagogy) and styles of interactions with students (interpersonal relationships) in use by teachers. The core is personal and may be perceived by the teacher as an outgrowth of his or her personalities and ethos, based on particular talents or skill sets that are unique to the person, and/or based on teaching experience. It may reflect what one believes about the nature of teaching and learning or about the developmental needs of their students, though these beliefs may not be articulated explicitly.

ever, some features of PFP, such as the increased focus on student achievement, use of baseline data and more effective uses of data at the end of the year (to see how they have done), reside in Level One and have become part of core practice at least for the duration of the pilot. One group insists that increased focus on student achievement is a result of *CSAP,* not PFP, and that the school was "already doing all of this, anyway." Most of the teachers believe that they will continue to use what they have learned from the pilot even if it does not become institutionalized or if additional compensation does not continue. Several teachers confessed that now that they were used to the process and knew how to write a successful objective, they would just as soon that Pay for Performance continued so they got the bonuses.

Study participants said that district or state mandates influence their core practice even when they do not agree philosophically with the change. As the study was underway, teachers in three of the schools were implementing a new district literacy program. They used this program as an example of a change in practice because it is mandated. Early on, they thought that the program, which was prescriptive, was making them worse teachers, but as the year wore on, many felt less negative about the program and more certain that they could make the required changes.

The mandated part of PFP (e.g., what each school agreed to as part of joining the pilot and becoming eligible for bonuses) is to write two objectives for student achievement based on baseline data (prior year usually), designate the expected attainment or growth, and measure it at the end of the year for additional compensation. Based on the reports of most of the teachers in the study, these objective setting and measuring activities have become part of their core practice. But most do not

believe that they have fundamentally changed the way that they teach as a result of PFP.

When asked what they believed the district and the Association expected to happen when the pilot was initiated, they said that the district wanted higher test scores and more accountability from teachers and that the Association was supporting the pilot in order to find out what a pay for performance or merit system is all about. Some teachers in the study observed that those teachers who are not philosophically in favor of performance-based compensation may be unwilling to attribute positive results or changes in their teaching to the pilot even where they actually exist.

To answer the key question of the study, the work with the 16 teachers showed that the apparent inconsistency in the data—changes in focus, use of baseline data, and awareness of student growth versus no changes in classroom practice attributable to the pilot—is a function of how teachers compartmentalize and separate their planning and assessment functions from "teaching activity" in the classrooms when thinking and talking about their practices. The product of their explanations of how they conceived change in practice is captured in *Figure 5-5*.

The qualitative study, like the objective-focused interview study, underlines the significance of a teaching staff's engagement with the building principal and with one another in pursuit of the best teaching practices on behalf of students. One teacher summed up what had been a theme in all of the focus group sessions: "I would rather any day to have good leadership, professional development, and time with my colleagues than $1500." Professional development was best for these teachers when they could learn with their building colleagues and principals. Some teachers in the study felt limited by what their students brought to the classroom; others felt empowered by the challenge, but most longed for better professional development and collegial interactions that would assist them with difficult-to-teach students. It is a reminder that teachers cannot change when they do not learn new or better practices to adopt.[2]

D. The Credibility and Fairness of Objectives for Compensation Purposes

In the mid-point report, several concerns related to the fairness of using objectives as the basis of compensation were identified from teacher interviews and surveys: (1) the diversity of students; (2) the potential for other teachers setting lower expectations or otherwise manipulating the data; (3) the fallacy of judging student outcomes on one measure; (4) the issue of judging teachers based on what their students do; and (5) the use of standardized tests as a measure of student performance. Other fairness issues not directly related to objectives included the potential for principal bias or gaps in skills and knowledge to influence outcomes; and the variation in the way special subject teachers, and specialists (non-classroom teachers and other service providers) were developing and assessing objectives in the elementary schools.

In the last two years of the pilot, while several issues continued to be of high importance to teachers, specific efforts and pilot learnings removed some of the concerns. For example, a differentiated rubric was developed for specials and specialists. Secondly, teachers came to understand that student diversity is controlled for in the process by (1) using baseline data and (2) allowing expected growth percentages less than 100. The following chart (*Figure 5-6*) shows the results from the Spring 2003 survey where teachers were asked to indicate the conditions or supports that would increase their confidence in a compensation plan based, in part, on student achievement.

In 2003, pilot teachers considered the most important condition of a compensation plan to be that "each student's growth is measured from his/her starting point at the beginning of the year" (93%). This same criterion was also ranked first (95%) in the 2002 survey. The second ranked condition is that "teachers are able to set objectives for student achievement in their class" (91%), indicating that school or district-established objectives were less desirable than teacher-developed objectives. The third ranked criterion for both 2002 and 2003 is "alternate measures of student achievement for use by specials, specialists,

FIG. 5-6

Conditions and Supports for Confidence in Compensation Plan, Spring 2003

Compensation Conditions	Strongly Agree/ Agree	Strongly Disagree/ Disagree	N	Rank*
District standards, instruction, and assessments are aligned.	84.9%	15.1%	357	4
Teachers are able to set objectives for student achievement in their class	91.0%	9.0%	366	2
Each student's growth is measured from his/her starting point at the beginning of the year	93.1%	6.9%	363	1
Teachers use district-approved assessments that are appropriate for their grade level and subject matter	72.2%	27.8%	360	8
Assessments administered using consistent and standardized procedures across the district	72.6%	27.4%	361	7
The relationship between the formal teacher evaluation system and a compensation plan is clearly defined	77.0%	23.0%	357	6
One measure is used to gauge student achievement in the compensation plan	37.1%	62.9%	356	9
Two or more measures are used to gauge student achievement in the compensation plan	78.0%	22.0%	355	5
Alternate measures of student achievement are available for use by specials, specialists, and special educators	89.7%	10.3%	360	3
Compensation Supports				
The principal has the skills and knowledge to administer a compensation plan fairly	80.2%	19.8%	364	5
Professional development addresses the learning needs of students in my school	86.3%	13.7%	364	2
Professional development addresses the appropriate administration of assessments	77.3%	22.7%	362	6
Professional development addresses the appropriate use of assessments	81.1%	18.9%	360	4
Teachers have ready access to student achievement data	89.9%	10.1%	366	1
An opportunity to work on student achievement issues with colleagues	86.2%	13.8%	363	3
Parent knowledge of and support for the compensation plan	64.2%	35.8%	355	7

*Based on percent strongly agree/agree

and special educators" (93% in 2002; 90% in 2003). In both 2002 and 2003, an important condition was the use of two or more measures of student achievement (93% in 2002 and 78% in 2003). In Spring 2003, this was further confirmed when only 37% of the respondents supported the use of one measure to gauge student achievement; however the steady decline of this condition may indicate that many pilot teachers had become comfortable with a single measure.

The most important support identified in the 2003 survey was ready access to student achievement data by teachers (90%). Respondents also identified the need for professional development that addresses the learning needs of the students in their school (86%) and an opportunity to work on student achievement issues with colleagues (86%).

From the early years of the pilot to the later years, teacher participants have gained experience and sophistication with the concept of pay for performance. Although the items in the last survey are ranked by the largest percentage of strongly agree/agree, they are mostly all of high importance to people who have been involved with the pilot.

Interview data for this same year (2003) indicate that pilot teachers as well as some control school teachers and central administrators are still concerned about (1) the potential for individuals gaming or rigging the system; (2) the potential for principal bias in the signoff of the objectives; (3) the inequity between classroom teachers and specials and specialists; and (4) issues of inconsistent administration from school to school (i.e., that some principals are more lenient than others).

Surprisingly, the emphasis on principal bias, identified earlier in the pilot, has tilted somewhat toward the bias that teachers may have on their own behalf:

> "Assessments—how can you grade your own work?"–Pilot teacher

> "I think there is always going to be a way for teachers to try to beat it or cheat it (PFP) which will affect the results."–Pilot teacher

> "The process (PFP) is flawed—teachers write their own objectives, do their own testing, may even make up their own tests, do their own scoring. There are differences in the grading/scoring. Even with an open rubric, it requires teacher judgment. No one checks the scoring—just look at the outcomes. It's bad because they create their own objectives and decide whom to exclude."–Pilot teacher

> "I would like to see guidelines changed so that teachers can't set a target lower than 80%."–Pilot principal

> "I could see a danger in what goals are set and the care or lack of care that some principals might take in checking the appropriateness of the goals."–Pilot teacher

Interviewees had many suggestions for improving Pay for Performance, indicating that for most, it is workable but revisions are still needed. Many of their recommendations are discussed in Chapters VII and VIII. Some examples follow:

> "I think that with PFP we've gotten a lot of information on how to write objectives and how to use OASIS. But it's missing the human element. How can we make this more successful? How can we hone in on making students achieve? Maybe focusing more training on classroom skills and management … how they [teachers] can operationalize their objectives, giving teachers tools."–Pilot teacher

> "The principal signs off on whether the objectives were met or not met. If the teacher is not on the same page as the principal, this could be a problem. Maybe there needs to be a committee of staff, colleagues, principal, and a community person."–Parent

> "There isn't much equity among objectives. I would like to see the bar raised on objectives. Objectives can be changed mid-course which allows people to lower the bar if kids aren't doing well."–Pilot teacher

E. Pilot Principals and Teacher Evidence of Attainment

As explained earlier in this report, Pay for Performance relies on an objective setting process that includes the concurrence of building principals, who then, review the body of evidence or data provided by a teacher to confirm that he or she has met the objectives. As indicated above, there

is concern among teachers that there is an inconsistency in the way principals carry out their obligations in this process. In the Spring 2003 survey data (see *Figure 5-6*), more than 80% of pilot teachers agreed or strongly agreed that a principal with the skills and knowledge to administer a compensation plan fairly is important to their confidence in a compensation plan that is based, in part, on student achievement. Additionally, teacher and principal interview data indicate concerns about the variation in the process by which principals review the "bodies of evidence."

In order to gain a picture of how principals respond, five sets of evidence presented by teachers to principals were gathered from a sample of five elementary schools. The samples were reviewed for eight factors. *Figure 5-7* shows the presence of the following eight factors related to the data collected.

a. Individual assessment scores for students for each classroom, which are important because expected growth is based on the number or percent reaching the target and not the class performance average
b. Pre-test scores for each student
c. Post-test scores for each student
d. Use of the Design Team Reporting Form, which, if used, will show that teachers have not reported averages
e. Supporting data, if needed
f. Objectives attached
g. Presence of the principal's organizational system for PFP
h. Evidence that PFP records are maintained from year to year

As *Figure 5-7* demonstrates, how principals are reviewing evidence to determine whether teachers have or have not met their objectives appears inconsistent among the five schools reviewed. The principal of School One has a system for collecting the objectives, a cover sheet for listing teachers and met/not met status, and a system for recording and following teacher attainment for the purpose of giving feedback and suggestions; the principal of School Three has a consistent process and record as well. The principals of Schools Two and Four do not have a consistent method or record, and the principal of School Five "does not look for data at the beginning of the year because teachers know how to use data and are trusted to set appropriate objectives" and does not believe it necessary to keep any of that information.

The concerns of both teachers and principals about the consistency of effort may be well founded. The fact that principals show inconsistent efforts and skill in working with data that teachers present at the end of the year leads to a perception of unfairness whether it exists or not, but mostly, where the principal's process is not thorough, opportunities for that principal to exert a positive influence on student achievement and

FIG. 5-7

Bodies of Evidence in Five Schools, 2003

Desirable Attributes of Bodies of Evidence Reviewed by Building Principals	School One	School Two	School Three	School Four	School Five
Individual assessment scores for each classroom	Yes	Partial	Yes	Mixed	No
Pre-test scores for each student	Yes	Yes	Yes	Mixed	No
Post-test scores for each student	Yes	Yes	Yes	Mixed	No
Design Team Reporting Form used	Yes	No	No	Mixed	No
Supporting data (if needed)	Yes	No	Yes	No	No
Objectives attached	Yes	No	Yes	Mixed	No
Principal has organizational system for PFP records	Yes	No	Yes	No	No
Records are maintained from year to year	Yes	No	Yes	No	No

teacher growth through the PFP process may be lost. Principal interviewees point to several factors that may have led to uneven principal motivation and processes with the PFP evidence: (1) turnover; (2) inadequate communication from the district about the pilot; (3) lack of professional development in how PFP aligns with the role of the principal in improving student achievement; (4) a feeling on the part of principals that this work was heaped on an already full plate without their input; (5) a new supervisory structure for principals; and (6) dissatisfaction with a merit pay system that had been implemented for principals prior to PFP.

Nonetheless, principals at the pilot schools and control schools are identifying the potential of a good objective setting process in maintaining a focus on student achievement and fostering a dialogue among teachers about student growth. Recommendations for professional development for principals are contained in Chapter IX.

F. Summary

Looking at Pay for Performance from the perspective of the pilot school teacher, one can see that the outcomes of the pilot have been greatly influenced by: (1) the manner in which teachers were invited or persuaded to join a pilot for the study of a compensation plan linked to student achievement; (2) the skill, commitment, and accountability that each teacher brought to the implementation; (3) the lack of district direction, assessments, and staff development supports; (4) the helpful role of teacher collegial structures; and (5) the skill, commitment, stability, and accountability of the building principals.

Nonetheless, the learning for teachers in the pilot schools has been exceptional. Regardless of what happens to Pay for Performance, pilot teachers have learned about and can talk about objectives in such a way that it is unlikely that they will return to previous objective-writing practices. Even among the lesser skilled and naysayers, the concept of a well-scaffolded objective has caught on; among the more skilled and open teachers, a well-scaffolded objective has become the precursor to greater student growth. The ability to craft objectives is improving.

There are clearly teachers who believe that Pay for Performance will work with some needed improvements to the instructional delivery system and the quality of school leadership, both of which should be strengthened whether or not there is Pay for Performance in the district.

CHAPTER VI

Quantitative Analyses

A. Introduction

The fundamental measurement of student achievement in the design of Pay for Performance is at the classroom level: the classroom teachers establish objectives based on an assessment of their own choosing and measure student attainment at the end of the year. Inherent in this model are the personal and professional judgments of the teacher and principal which represent a depth of understanding about teaching and learning with a particular group of students. However, these individual classroom level results are not comparable within or across schools. For this reason, the study design includes an independent analysis of student performance on two standard measures, the *Iowa Test of Basic Skills (ITBS)* and the *Colorado Student Assessment Program (CSAP)*. These standardized tests provide a consistent measure with which to compare student achievement before and after the pilot and between pilot and control schools.

This chapter describes the analysis of student performance conducted by CTAC on these two standard measures. Measuring student achievement on standard measures has its own limitations as well. First, the Denver Public School curriculum is not clearly aligned to either test. Secondly, the pilot coincided with a time in the district when these assessments were in flux. The *CSAP*, a new assessment for the state of Colorado, was phased in over the life of the pilot. Although designed to be a criterion-referenced test, the *CSAP* has been scaled to allow for year-to-year comparisons of individual students. There has been a de facto phasing out of the *ITBS* in the district coincident with the first year of the pilot when the state test gained more importance in the district and schools. Finally, the district administration of the assessments does not involve clear criteria for student exclusions and allows principal discretion in exempting students from

the assessment. Thus, the analyses indicate not only variations in practice across schools, but also non-random exclusions. In particular, students who are not English proficient are more likely to be excluded. The assessments with their changes and limitations are discussed in Chapter III.

In order to analyze the effects of the pilot on student achievement, CTAC uses two statistical methods—Hierarchical Linear Modeling (HLM) and Individual Growth Modeling (IGM), which are explained in Chapter III and will be elaborated upon in this chapter. Teachers, in setting growth targets and reviewing the evidence of attainment at the end of the year, are able to take into consideration each student's past performance along with features of his or her current behavior and performance. This can be seen in the conditions and expectations set forth in objectives. For example, the attendance record of a student or his or her level of language acquisition may be considered by the teacher in setting expectations. CTAC does not have access to all of these factors under consideration when attainment of teacher objectives is reviewed. However, the HLM and IGM models are able to control for differences in school and student characteristics which are known to affect student achievement.

B. Design Features

Outcome Measure: Normal Curve Equivalent Score

The quantitative study design employs the use of normal curve equivalent (NCE) scores rather than scale scores. The NCE indicates where a student ranks relative to a reference population of other students in the same grade on a normal distribution curve. A difference of zero between this year's NCE score and last year's NCE score means that a student has achieved the academic growth expected for one year of development and instruction. This property of the NCE makes it possible to interpret a positive slope, or an increase in score over time, to mean that the student is performing better than expected based on previous scores—or attaining more than an expected year of growth. Conversely, a negative slope indicates that the student is not achieving as expected relative to the reference population, given that student's past performance. It is important to note that NCEs are not grade equivalents and that a statistically significant increase in NCEs from one year to the next is not a measure of the number of academic years increased.

Choice of ITBS and CSAP; Need for Weighting

Administration of the *Iowa Test of Basic Skills (ITBS)* was mandatory for all schools at the time the study began. During the study period, the *Colorado Student Assessment Program (CSAP)* was phased into use in all schools. The *CSAP*, although not originally designed to make comparisons across years, has now been scaled, allowing year-to-year comparisons for the same student. CTAC converted the scale scores into normal curve equivalents, using the local school district as the reference population. The *ITBS* normal curve equivalents are referenced to a national population. The *ITBS* and *CSAP* assessments each have three components: the *ITBS* has Reading, Language, and Math tests, while *CSAP* has Reading, Writing, and Math tests. During the phase-in period the *CSAP* was not administered to every grade, and the individual component tests were not all administered to the same grades. This means that for analysis purposes, two consecutive years of test scores are not always available.

The *ITBS* is given in the fall and spring. This analysis uses only spring scores because testing in the fall is unusual in the Denver schools.

Unfortunately for the pilot's purposes, *ITBS* testing became voluntary rather than mandatory in the district during the first year of the pilot and testing rates fell dramatically in both pilot and control schools. Although *CSAP* is state mandated, testing rates for *CSAP* also differed across schools and grades. An analysis of testing rates by student demographic factors showed that testing was not random for both *ITBS* and *CSAP*. For example, *Figure 6-1* shows that with the exception of high schools in 2001, the *ITBS* and *CSAP* reading tests were least likely to be administered to non English proficient students, while for the most part bilingual students were more likely to take the tests than native English speakers. Variation at the school

FIG. 6-1

ITBS and *CSAP* Reading Testing Rates by Grade, English Proficiency and SES 2001-2002 School Year

		English Proficiency			SES		
Level/Grade	**Overall**	**Not Proficient**	**Bilingual**	**English Only**	**Higher**	**Lower**	**N**
***ITBS* Reading Testing Rates (Percent)**							
Elementary							
2	76	50	85	89	92	71	3049
3	83	62	90	90	94	80	2926
4	85	65	92	90	94	82	3078
5	86	66	91	91	95	83	3078
6	92	92	88	93	100	90	49
Middle							
6	69	38	86	82	83	69	926
7	66	39	84	74	68	66	842
8	68	44	85	74	81	67	781
High							
9	63	28	65	70	74	56	2699
10	52	23	56	55	53	51	1800
11	50	29	54	51	53	47	1366
***CSAP* Reading Testing Rates (Percent)**							
Elementary							
3	92	90	93	92	95	91	2926
4	91	86	93	93	95	90	3078
5	88	70	93	92	95	86	3078
6	90	83	88	93	100	88	49
Middle							
6	77	54	90	86	80	77	926
7	77	56	90	85	82	77	842
8	77	57	90	83	86	76	781
High							
9	76	60	78	79	82	72	2699
10	67	45	69	70	74	61	1800

level (not shown here) is greater. Exemptions from standardized testing for students with disabilities or students who are not English proficient, are at the discretion of each school. Rather than excluding non-English speakers and students with disabilities from the analysis, we chose to weight the data to reduce the possibility that pilot effects are due to differences in testing policy between pilot and control schools.

In particular, low SES (socioeconomic status based on student participation in the free and reduced lunch program) and non-English proficient students were tested at lower rates across schools, thus the data have been weighted to reflect the population distribution of SES and English proficiency within year, school, and grade. The rate of testing also differs within standardized test—students are less likely to take the Math, Language and Writing components than the Reading component. Thus, six weights were developed, one for each test component. By weighting, we reduce the possibility that a difference in achievement is attributable to differences in testing policies between schools, rather than due to the pilot. The results of the weighting process for the *ITBS* Reading test sample for the baseline year are found in *Figure 6-2.* Looking at the elementary schools, we see that the actual percentage of students who are not English proficient is 25% for the pilots and 21% for the controls, based on the students present in October. In the sample of students who were tested in the baseline year, non-English proficient students are under represented —20% of the pilot group and 15% of the control group. In the weighted sample, the distribution is closer to the actual population distribution—24% of the weighted pilot students and 20% of the weighted controls are non-proficient. The weighted sample does not precisely reflect the October school population since some of the proficiency/SES groups within schools had no students tested.

Pre/Post—Pilot/Control Comparisons

As mentioned earlier in this chapter, the slope (or change) in NCEs over time have the useful property of measuring whether students have attained less than a year's expected growth (a statistically significant negative slope), a year's growth (a slope which is not significantly different from zero), or more than a year's growth (a statistically significant positive slope). To measure the effect of Pay for Performance, the analysis compares the average of the slopes of pilot school students to the average of the slopes for control students. A positive and statistically significant slope for the pilot students indicates that pilot students attained more than a year's growth. However in order to assess whether this increase would have happened without the pilot treatment, we also compare the pilot slope to the slope for the control students. If the control slope is equal to or higher than the pilot slope, we conclude that the pilot treatment has not increased student achievement—or that the increase seen in pilot student scores would have happened without the pilot treatment.

Thus it is the difference between pilot and control slopes in the current analysis which measures the effect of the pilot. A positive difference which is statistically different from zero demonstrates that the pilot had a positive effect on student achievement over the course of the study period. Similarly, a negative difference in slopes demonstrates that the pilot had a negative effect on student achievement. A result that is not statistically different from zero demonstrates that the pilot had no effect on student achievement.

Pilot School Selection

The Design Team presented the PFP pilot to most of the elementary schools, and invited all of the middle and high schools to join the pilot. Following a presentation, the teachers voted on whether or not to participate. Schools at which 85% (later 67%) or more of the teachers voted affirmatively became pilot schools. A full list of pilot and control schools is found in Chapter III. Manual High School underwent a major reorganization during the first year of the pilot, so separate analyses were performed for the two pilot high schools, because the effect of the reorganization of Manual cannot be separated from the effect of the pilot.

Allowing schools to self-select has the advantage of gaining the cooperation of teachers, but it also poses a threat to the validity of the research. There may have been an unmeasured (or impossible to measure) 'latent' characteristic that caused some schools to select into the pilot. For example, one

FIG. 6-2

Demographics for *ITBS* Reading Sample, Baseline Year Unweighted and Weighted, by Level of School

	October Count				Unweighted Sample				Weighted Sample			
	pilot		control		pilot		control		pilot		control	
	n	%	n	%	n	%	n	%	n	%	n	%
Elementary Schools												
Not English Proficient	948	25	1980	21	614	20	1118	15	690	24	1352	20
Bilingual	585	15	1336	14	500	16	1141	15	430	15	1035	15
English Only	2327	60	6037	65	1942	64	5300	70	1720	61	4408	65
Any Disability	592	15	1379	15	462	15	1106	15	446	16	969	14
No Disability	3268	85	7974	85	2594	85	6453	85	2394	84	5826	86
Male	1950	51	4820	52	1542	50	3900	52	1409	50	3563	52
Female	1910	49	4533	48	1514	50	3659	48	1431	50	3232	48
Lower SES	2762	72	6952	74	2121	69	5448	72	2030	71	5066	75
Higher SES	1098	28	2401	26	935	31	2111	28	810	29	1729	25
Native American	34	1	88	1	31	1	75	1	27	1	61	1
Black	599	16	1392	15	526	17	1235	16	478	17	1077	16
Asian	94	2	226	2	79	3	204	3	67	2	176	3
Hispanic	1406	36	3491	37	1080	35	2701	36	1046	37	2547	37
White	1727	45	4156	44	1340	44	3344	44	1222	43	2933	43
Middle Schools												
Not English Proficient	469	30	2078	16	199	21	1294	13	454	30	2057	16
Bilingual	405	26	2272	18	306	32	1926	19	391	26	2232	18
English Only	666	43	8327	66	453	47	6915	68	666	44	8254	66
Any Disability	182	12	1531	12	122	13	1034	10	177	12	1307	10
No Disability	1358	88	11146	88	836	87	9101	90	1334	88	11236	90
Male	780	51	6496	51	488	51	5115	50	748	49	6386	51
Female	760	49	6181	49	470	49	5020	50	763	51	6157	49
Lower SES	1364	89	7196	57	885	92	5772	57	1364	90	7173	57
Higher SES	176	11	5481	43	73	8	4363	43	147	10	5370	43
Native American	16	1	103	1	14	1	93	1	37	2	123	1
Black	28	2	2302	18	23	2	2043	20	86	6	2463	20
Asian	7	0.5	343	3	6	1	336	3	7	0.5	392	3
Hispanic	989	64	4424	35	674	40	3555	35	995	66	4502	36
White	500	32	5505	43	241	25	4108	41	386	26	5062	40

FIG. 6-2 CONTINUED

Demographics for *ITBS* Reading Sample, Baseline Year Unweighted and Weighted, by Level of School

	October Count				Unweighted Sample				Weighted Sample			
	pilot		control		pilot		control		pilot		control	
	n	%	n	%	n	%	n	%	n	%	n	%
High Schools												
Not English Proficient	249	14	1316	12	129	12	354	12	238	13	1156	14
Bilingual	315	18	2383	22	211	19	773	27	315	18	1814	22
English Only	1214	68	7133	66	760	69	1732	61	1214	69	5391	64
Any Disability	254	14	1183	11	117	11	235	8	199	11	509	6
No Disability	1524	86	9649	89	983	89	2624	92	1568	89	7853	94
Male	883	50	5552	51	514	47	1417	50	824	47	4341	52
Female	895	50	5280	49	586	53	1442	50	943	53	4019	48
Lower SES	994	56	5689	53	578	53	1675	59	992	56	4597	55
Higher SES	784	44	5143	47	522	47	1184	41	775	44	3764	45
Native American	5	0.3	62	0.6	4	0.4	28	1	7	0.4	121	1
Black	288	16	1189	11	212	19	520	18	372	21	1414	17
Asian	12	0.7	261	2	8	0.7	122	4	13	0.7	246	3
Hispanic	343	19	2566	24	245	22	1018	36	402	23	2038	24
White	1130	64	6754	62	631	57	1171	41	974	55	4543	54

could hypothesize that pilot schools have leadership that is willing to take chances while the control schools have leadership that is conservative about change. The analyses cannot rule out that differences in achievement between control and pilot schools are due to this latent characteristic. Willingness to participate in a research study may be related to the overall achievement level of a school—teachers of high achieving students may be more willing to be scrutinized than teachers of low achieving students. Since we are looking for growth in average NCE scores, we are more likely to see gains among the lower achieving students. If schools which have higher average achievement are more likely to self-select into the pilot, the results would be biased against seeing a pilot effect. This may indeed have happened at the elementary level where the baseline average *ITBS* Reading score is 43.8 for elementary pilot schools and 40.8 for the control schools. Starting with pilot schools which have higher achievement levels at baseline may also introduce bias due to regression to the mean, this form of bias would make it more likely to see a negative effect. In contrast, the middle school pilots had lower average *ITBS* Reading scores at baseline—32.6 for pilot schools versus 42.9 for control schools, here the bias is towards finding a positive effect. Two schools of very different baseline achievement levels participated as pilots at the high school level. Manual had a mean *ITBS* Reading NCE of 34.8, Thomas Jefferson had a mean of 55.8, and the control schools averaged 44.2 NCEs. Thus Manual is biased toward a positive effect while Thomas Jefferson is biased towards a negative effect.

Self selection also restricts the representativeness of the pilot sample, making the results only applicable to PFP programs which are instituted with teacher approval.

Control Schools and Treatment Contamination

For the elementary school analysis, three control schools were chosen by the district as matches for each pilot school on the basis of demographic characteristics as described in Chapter III. At the secondary level, all of the non-pilot middle schools serve as middle school controls and all of the non-pilot high schools serve as high school controls. Generally, matching introduces bias into the analysis and imposes limits on the amount of information which the matching characteristics can provide. For instance, matching on socioeconomic characteristics at the elementary level precludes us from analyzing the influence of SES, English proficiency, and school enrollment on the pilot outcomes.

Half of the elementary school controls and all of the middle and high school controls were selected from schools that had been recruited for the pilot and voted not to participate. Controls by definition should not be aware of the pilot. This introduces another source of bias that would tend to dilute the effect of the pilot program. There is anecdotal evidence to indicate that several control school principals and teachers implemented the PFP objective writing process or a modified version of it. When control schools implement portions of the treatment, the contamination of the study design makes it more difficult to detect an effect of the pilot program on student achievement. Secondly, during the 2002-2003 school year, most of the elementary schools, including all of the pilot elementary schools but one, took part in a literacy initiative which required teachers to write literacy objectives. Thus for the last year of the study, part of the pilot's "unique" treatment occurred in both pilot and control elementary schools. This will bias the results toward observing no pilot effect on reading tests.

School, Student, and Teacher Characteristics

Factors other than the pilot 'treatment' affect student achievement and these factors differ between schools. To insure that the estimates of pilot effectiveness are not due to differences in school populations certain school and student characteristics have been controlled for and where possible teacher characteristics as well.

The school characteristics used in the student achievement analysis are taken from the Denver Public Schools report cards from school years 1998-1999 through 2000-2003. CTAC chose a subset of the reported measures that were available for the whole study period. To control for the lack of continuity in school administration, we included the number of years the principal has been at the school. Factors that control for differences in student population include the following: percent of students with low SES (measured by participation in the free/reduced lunch program); percent of students with a disability; and percent of students classified as English language learners. The percent of teachers not fully licensed is used to control for qualitative differences in the teacher population between schools. School enrollment provides a control for overall size of the school. All of these school factors, with the exception of principal years at the school, have been centered at the grand mean by type of school: for elementary schools the mean of the pilot and control schools participating in PFP was used; for middle schools the mean for all middle schools in Denver was used; and for the high schools the grand mean is based on the mean for all high schools. Centering the school characteristics makes it possible to interpret their coefficients as an increase of one unit above the average Denver Public School at the middle and high school levels. At the elementary school level, this equates to the average elementary school participating as either a pilot or control.

The student demographic data collected by the Denver Public Schools provide measures of the non-school influences on a student's performance. For this study, gender, ethnicity, language proficiency, grade, the presence of any disability, participation in the free/reduced lunch program, and grade retention were collected.

Teacher characteristics are available for the study years, but not the baseline years, and for the elementary and middle schools, the teacher characteristics are not available for the control schools for the first pilot year. This lack of data makes it impossible to use teacher characteristics in the pre/post pilot/control analyses. In the analyses of the post period that include only pilot schools, teacher characteristics are used. The characteristics collected from the DPS Human Resource files are

degree (bachelor's, master's, or doctorate degree), years of experience in the DPS system, and whether the teacher is part of the Teacher-in-Residence (TIR) program, an alternative certification program. Class lists linking teachers to students were collected and entered by hand for the 1999-2000 school year for the elementary pilot schools. After that, records were obtained from the DPS electronic files at three points during the school year. These class lists were used to link teacher data to student achievement records. After the teacher and student data were linked, an indicator was created to identify students who had two or more teachers during one school year.

C. Comparison of Pilot Schools to Control Schools

Student achievement scores are not independent observations. Students are grouped within classrooms and schools, and at each level of the hierarchy the student's scores are correlated. An individual student's scores are also correlated across years. Two-stage hierarchical linear modeling (HLM) is used in this analysis to appropriately control for the lack of independence among observations. The HLM analysis groups children within schools only because classroom assignment data are not available for all years. Collection of teacher assignments was not done retroactively, thus, we do not know classroom assignments for students during the baseline year.

In the first stage of the model, we predict a student's NCE score as the sum of an intercept for the student's school (α_{0j}), the effects of the pilot, time, the interaction of pilot and time, language proficiency, disability, ethnicity, gender, grade, and SES, and the random error (r_{ij}) associated with the *i*th student at the *j*th school. This model gives us intercepts for pilot ($\alpha_j + \beta 1$) and control schools (α_j), which tell us how the two groups compared before the pilot treatment began. It also estimates the control students' slope $\beta 2$, describing the change in student achievement scores over time for the control schools. The coefficient for the interaction of pilot treatment with time ($\beta 3$) measures whether the pilot schools have the same slope as the control schools. A positive and statistically significant value for $\beta 3$ indicates that PFP has improved student achievement and conversely a negative and statistically significant $\beta 3$ indicates that the pilot is associated with a decrease in student achievement. The pilot students' slope is calculated by adding together $\beta 2$ and $\beta 3$. A slope which is not significantly different from zero indicates one year of expected growth, a statistically significant

FIG. 6-3

PFP Effect—Elementary Schools—*ITBS* Weighted Two-Stage Hierarchical Linear Models

	***ITBS* Reading**			***ITBS* Language**			***ITBS* Math**		
	Unadjusted	Adjusted for School Factors	Adjusted for School and Student Factors	Unadjusted	Adjusted for School Factors	Adjusted for School and Student Factors	Unadjusted	Adjusted for School Factors	Adjusted for School and Student Factors
Control Intercept	43.5***	42.3***	56.1***	43.7***	43.5***	56.5***	43.6***	42.7***	52.6***
Pilot Intercept	43.8***	44.5***	58.6***	41.4***	42.7***	55.9***	45.9***	46.0***	56.4***
Difference Between Pilot and Control Intercepts	0.4	2.1*	2.5**	-2.3	-0.7	-0.6	2.3	3.3***	3.7***
Control Slope	-0.3***	-0.1	0.1	-0.2*	-0.1	-0.03	-0.6***	-0.4***	-0.3***
Pilot Slope	-0.1	-0.2	-0.2	-0.3*	-0.2	-0.1	-0.7***	-0.6***	-0.7***
Difference Between Pilot and Control Slopes	0.2	-0.1	-0.3*	-0.1	-0.1	-0.1	-0.1	-0.2	-0.4*

* statistically significant at $p < 0.05$; ** statistically significant at $p < 0.01$; *** statistically significant at $p < 0.001$

positive slope indicates more than an expected year of growth, and a statistically significant negative slope indicates less than a year's growth.

Level 1:

$$\gamma_{ij} = \alpha_j + \beta1(\text{Pilot}_{ij}) + \beta2(\text{Time}_{ij}) + \beta3(\text{Pilot}_{ij} \times \text{Time}_{ij}) + \beta4(\text{SES}_{ij}) + \beta5(\text{Disabled}_{ij}) + \beta6(\text{Retained a Grade}_{ij}) + \beta7(\text{Not Proficient}_{ij}) + \beta8(\text{Bilingual}_{ij}) + \beta9(\text{Native American}_{ij}) + \beta10(\text{Black}_{ij}) + \beta11(\text{Asian}_{ij}) + \beta12(\text{Hispanic}_{ij}) + \beta13(\text{Male}_{ij}) + r_{ij}$$

where $r_{ij} \sim N(0,\sigma^2)$

The Level 2 model expresses the intercept of school *j* as the grand mean and deviations from that mean associated with school level characteristics and a random error term (ϵ_{0j}).

Level 2:

$$\alpha j = \gamma + \beta14(\text{Principal Years at School}_j) + \beta15(\text{Percent Disabled}_j) + \beta16(\text{Percent English Language Learners}_j) + \beta17(\text{Percent Free/Reduced Lunch}_j) + \beta18(\text{Percent Teachers not Fully Licensed}_j) + \beta19(\text{Total Enrollment}_j) + \beta_j$$

where $\epsilon_j \sim N(0, \tau_{00})$

Three models are presented for each of the six tests, a model testing for pilot effect without adjusting for any covariates (Model A), a second model adjusting for school level covariates (Model B), and a third model adjusting for school and student level factors as described in the equations above (Model C). Four analyses were performed, one each at the elementary and middle school level, and two at the high school level. The full HLM models are presented in the Appendix, summary tables showing the estimated intercepts and slopes are presented in this chapter.

Elementary School PFP Outcomes

The unadjusted model (Model A) reflects most closely what happened 'in the real world'. The model adjusting for school factors (Model B) allows us to estimate what the results of PFP would have been had the characteristics of the pilot and control schools been equal, and the third model (Model C), adjusting for school and student factors, estimates the effects of PFP had the student populations been the same. The pilot and control intercepts represent the average achievement level of pilot and control students before the study began. The intercept in the third model is higher than the previous two models because the influences of being poor, disabled, failing the previous grade, lacking proficiency in English, or being bilingual, male, and non-white (not already controlled by the matching process) have all been removed. Thus, the intercept of the third model is

FIG. 6-4

PFP Effect—Elementary Schools—*CSAP*
Weighted Two-Stage Hierarchical Linear Models

	CSAP **Reading**			***CSAP*** **Writing**			***CSAP*** **Math**		
	Unadjusted	**Adjusted for School Factors**	**Adjusted for School and Student Factors**	**Unadjusted**	**Adjusted for School Factors**	**Adjusted for School and Student Factors**	**Unadjusted**	**Adjusted for School Factors**	**Adjusted for School and Student Factors**
Control Intercept	51.6***	51.0***	65.3***	49.5***	49.4***	63.0***	51.0***	49.4***	61.2***
Pilot Intercept	54.6***	53.0***	67.9***	55.6***	52.6***	66.8***	56.5***	54.6***	67.4***
Difference Between Pilot and Control Intercepts	2.9	2.1*	2.6**	6.1***	3.2**	3.7***	5.5*	5.2***	6.2***
Control Slope	-0.2	-0.01	0.2	0.1	0.3**	0.50***	0.002	0.2	0.2
Pilot Slope	-0.5**	-0.5***	-0.5***	-0.3	-0.4	-0.3	-1.3***	-1.1***	-1.3***
Difference Between Pilot and Control Slopes	-0.3	-0.5**	-0.7***	-0.5*	-0.7**	-0.8***	-1.3***	-1.3***	-1.5***

* statistically significant at $p < 0.05$; ** statistically significant at $p < 0.01$; *** statistically significant at $p < 0.001$

analogous to the average achievement level of a non-disabled white female student who is a native English speaker.

The elementary school HLM models are presented in full in the Appendix, *Figures A-1* through *A-6* and summarized in *Figures 6-3* and *6-4*. The elementary pilot students had higher *ITBS* Reading achievement levels than the controls at baseline, adjusting for school characteristics (Model B). The control intercept was 42.3 and the pilot intercept was 44.4, a statistically significant difference of 2.1 NCEs ($p<.05$). The difference is larger (2.5 NCEs, $p<.01$) when student characteristics are added (Model C). Ideally, the school and student factors should reduce the difference between controls and pilots at baseline, making the two groups comparable. In this case, and for *ITBS* Math and all three *CSAP* tests, the baseline difference between pilots and controls persists after adjusting for school and student characteristics. This is an indication that selection bias is present.

In *Figure 6-3* the slopes for the control and pilot students is calculated from the results of the HLM models. As previously mentioned the difference between the pilot and control slopes estimates the effect of the pilot. Because the achievement scores have been transformed into NCEs, a slope of zero represents one year of expected growth in achievement levels, a slope less than zero represents less than a year of growth and a slope greater than zero more than a year's growth.

In the unadjusted model, we see that over the course of the study the control students showed a significant decrease in *ITBS* Reading of 0.3 NCEs per year ($p<.001$) on average. This effect is smaller and non significant when school and student factors are included in the model. The pilot students also have a negative unadjusted slope of -0.1 NCEs per year. Had the school and student demographics of the pilot and control schools been equal, we estimate that the slope for the controls would have been 0.1 while the slope for the pilots would have been -0.3. Thus, the PFP effect for *ITBS* Reading at the elementary level is a statistically significant ($p<.05$) decrease of 0.3 NCEs per year.

All three of the *ITBS* Language models *(Figure 6-3),* estimate a negative PFP effect (-0.1), which is not statistically different from zero. No PFP effect has been detected for elementary level *ITBS* Language achievement.

Both control and pilot school students experienced statistically significant decreases in *ITBS* Math achievement levels over the course of the study. Holding school and student factors constant, we estimate that the pilot slope was -0.7 ($p<.001$) and the control slope was -0.3 ($p<.001$). This

FIG. 6-5

PFP Effect—Middle Schools—*ITBS*
Weighted Two-Stage Hierarchical Linear Models

	***ITBS* Reading**			***ITBS* Language**			***ITBS* Math**		
	Unadjusted	Adjusted for School Factors	Adjusted for School and Student Factors	Unadjusted	Adjusted for School Factors	Adjusted for School and Student Factors	Unadjusted	Adjusted for School Factors	Adjusted for School and Student Factors
Control Intercept	43.1***	42.6***	55.6***	46.4***	46.1***	57.7***	42.1***	41.4***	51.4***
Pilot Intercept	32.9***	40.2***	52.7***	39.7***	42.7***	54.9***	34.9***	40.4***	49.9***
Difference Between Pilot and Control Intercepts	-10.3	-2.4	-2.9	-6.7	-3.3	-2.8	-7.2	-0.9	-1.5
Control Slope	-1.1***	-1.3***	-0.4***	-1.0***	-1.1***	-0.4***	-0.8***	-1.0***	-0.2*
Pilot Slope	-0.3	-0.4	0.7**	-1.2***	-1.5***	-0.7**	-0.5*	-0.6*	0.2
Difference Between Pilot and Control Slopes	0.8**	0.9**	1.1***	-0.3	-0.4	-0.3	0.3	0.4	0.3

* statistically significant at $p < 0.05$; ** statistically significant at $p < 0.01$; *** statistically significant at $p < 0.001$

FIG. 6-6

PFP Effect—Middle Schools—*CSAP*
Weighted Two-Stage Hierarchical Linear Models

	CSAP **Reading**			***CSAP*** **Writing**			***CSAP*** **Math**		
	Unadjusted	**Adjusted for School Factors**	**Adjusted for School and Student Factors**	**Unadjusted**	**Adjusted for School Factors**	**Adjusted for School and Student Factors**	**Unadjusted**	**Adjusted for School Factors**	**Adjusted for School and Student Factors**
Control Intercept	49.5***	49.0***	63.5***	49.7***	49.5***	64.2***	50.0***	49.6***	59.9***
Pilot Intercept	40.6***	49.5***	64.4***	40.7***	48.5***	63.2***	41.1***	48.7***	58.6***
Difference Between Pilot and Control Intercepts	-8.9	0.5	0.9	-9.0	-0.9	-1.1	-8.9	-0.9	-1.3
Control Slope	0.03	-0.3*	0.5***	0.003	-0.3**	0.6***	-0.3*	-0.6***	0.6***
Pilot Slope	0.3	-0.2	0.5	0.3	0.2	1.2***	1.6***	1.3***	2.2***
Difference Between Pilot and Control Slopes	0.3	0.1	-0.1	0.3	0.5	0.7*	1.9***	1.9***	1.6***

* statistically significant at p < 0.05; ** statistically significant at p < 0.01; *** statistically significant at p < 0.001

results in a statistically significant negative pilot effect of -0.4 NCEs per year (p<.05).

The elementary pilot students had baseline *CSAP* achievement levels which were 2.9 NCEs higher for Reading, 6.1 NCEs higher for Writing, and 5.5 NCEs higher for Math than the control students (*Figures 6-4* unadjusted models). These differences in intercepts persist after adjusting for school and student characteristics. Statistically negative PFP effects are estimated for *CSAP* Reading (-0.7, p<.001), Writing (-0.8, p<.001) and Math (-1.5, p<.001).

Middle School PFP Outcomes

At the middle school level, we compare the two pilot middle schools to all of the other Denver middle schools. The full descriptions of the HLM models are found in the Appendix, *Figures A-7* through *A-12*. The pilot middle schools at baseline have an *ITBS* Reading level below that of the controls, and the differential between pilots and controls is similar on each of the six tests (*Figures 6-5* and *6-6*). By controlling for school and student characteristics, we eliminate much of the difference in baseline achievement levels. For all six tests, the difference between the adjusted intercepts is smaller and not statistically significant.

The control students experienced statistically significant decreases in *ITBS* scores over the study period of 0.4 NCEs per year (p<.001) on the Reading and Language tests and 0.2 NCEs (p<.05) on the Math test. The pilot students performed significantly better than the control students on the *ITBS* Reading exam, with a slope of 0.7 (p<.01). This is a statistically significant PFP effect of 1.1 NCEs per year (p<0.001) more than the control students. On the Language exam, the middle school pilot students lost 0.7 (p<.01) NCEs per year on average. This represents a PFP effect of -0.3, which is not statistically different from zero. Pilot students showed a small and non-significant increase of 0.1 NCEs per year on the *ITBS* Math exam, 0.3 NCEs per year better than the controls.

On the *CSAP* exams, the control school students showed a statistically significant amount of improvement of 0.5 NCEs per year (p<.001) on the Reading test and 0.6 NCEs per year (p<.001) on the Writing and Math tests. The pilot school students performed about the same as the control school students on the Reading test as the PFP

FIG. 6-7

PFP Effect—High Schools—*ITBS*
Weighted Two-Stage Hierarchical Linear Models

	ITBS Reading			*ITBS* Language			*ITBS* Math		
	Unadjusted	Adjusted for School Factors	Adjusted for School and Student Factors	Unadjusted	Adjusted for School Factors	Adjusted for School and Student Factors	Unadjusted	Adjusted for School Factors	Adjusted for School and Student Factors
Control Intercept	45.2***	44.6***	55.5***	35.7***	34.1***	43.7***	40.7***	41.0***	50.8***
Manual HS Intercept	32.2***	46.8***	51.2***	21.8***	15.3**	15.8*	23.0***	28.2***	34.5***
Jefferson HS Intercept	55.9***	41.5***	53.1***	25.9**	20.8*	26.2*	25.6**	15.0*	24.7*
Difference Between Manual and Control Intercepts	-13.0	2.2	-4.3	-13.9**	-18.8**	-28.0**	-17.7**	-12.8*	-16.2*
Difference Between Jefferson and Control Intercepts	10.7	-3.1	-2.4	-9.7	-13.3	-17.5	-15.1	-26.0**	-26.1*
Control Slope	0.9***	1.4***	2.0***	2.2***	2.7***	3.5***	3.5***	3.8***	4.2***
Manual HS Slope	0.7	1.3	3.8***	7.2***	8.2***	11.7***	6.7***	7.7***	9.1***
Jefferson HS Slope	-1.0	-0.8	0.4	9.5***	10.2***	10.5***	16.4***	17.1***	16.8***
Difference Between Manual and Control Slopes	-0.2	-0.01	1.9	5.0***	5.5***	8.2***	3.2*	3.9**	4.8***
Difference Between Jefferson and Control Slopes	-2.0**	-2.1**	-1.6*	7.3**	7.5**	7.1**	12.9***	13.4***	12.6***

* statistically significant at p < 0.05; ** statistically significant at p < 0.01; *** statistically significant at p < 0.001

effect is -0.1 and not statistically different from zero. Pilot students outperformed control students by 0.7 NCEs per year (p<.05) on the Writing exam and 1.6 NCEs per year (p<.001) on the Math exam. Pilot school students increased an average of 1.2 NCEs (p<.001) per year on the Writing exam and 2.2 NCEs (p<.001) per year on the Math exam.

High School PFP Outcomes

The six high school HLM models may be found in *Figures A-13* through *A-18* in the Appendix. At baseline the unadjusted average *ITBS* Reading scores for Manual students were 13 NCEs below the control students and the Thomas Jefferson High students were 10.7 NCEs higher than the controls *(Figure 6-7).* After adjusting for school and student characteristics, the differences between the controls and the two pilot schools were smaller and not statistically significant. Similar results occurred for the *CSAP* test *(Figure 6-8);* however, for the *ITBS* Language and Math tests significant baseline differences still exist between pilot and control students in the adjusted model.

The control school students increased their *ITBS* Reading scores by 2 NCEs per year (p<.001) on average during the study, while Manual students increased at 3.8 NCEs per year (p<.001) and Thomas Jefferson students increased at a rate of 0.4 NCEs per year. The positive PFP effect of 1.9 for Manual is somewhat significant (p=0.09). The negative effect (-1.6, p<.05) for Thomas Jefferson students indicates that Thomas Jefferson students, despite the intervention, showed less growth than

FIG. 6-8

PFP Effect—High Schools—*CSAP*
Weighted Two-Stage Hierarchical Linear Models

	***CSAP* Reading**			***CSAP* Writing**			***CSAP* Math**		
	Unadjusted	**Adjusted for School Factors**	**Adjusted for School and Student Factors**	**Unadjusted**	**Adjusted for School Factors**	**Adjusted for School and Student Factors**	**Unadjusted**	**Adjusted for School Factors**	**Adjusted for School and Student Factors**
Control Intercept	50.5***	52.3***	65.3***	50.4***	51.0***	64.0***	51.6***	51.0***	60.3***
Manual HS Intercept	38.2***	51.5***	60.7***	34.8***	51.5***	57.6***	36.7***	51.5***	58.0***
Jefferson HS Intercept	56.6***	45.7***	55.6***	58.2***	51.4***	62.2***	57.2***	53.0***	60.4***
Difference Between Manual and Control Intercepts	-12.3	-0.8	-4.6	-15.6*	0.5	-6.4	-14.9*	0.5	-2.3
Difference Between Jefferson and Control Intercepts	6.0	-6.6	-9.8	7.8	0.3	-1.8	5.6	2.0	0.1
Control Slope	0.2	0.4*	1.2***	0.2	0.5	1.7***	0.1	0.3	1.3***
Manual HS Slope	0.5	0.2	2.1*	1.6	0.2	2.8*	1.3	-0.3	1.8
Jefferson HS Slope	0.5	0.9	2.6***	-0.2	-0.3	1.9**	0.2	-0.5	1.2
Difference Between Manual and Control Slopes	0.3	-0.2	1.0	1.3	-0.3	1.1	1.2	-0.5	0.5
Difference Between Jefferson and Control Slopes	0.2	0.5	1.5*	-0.5	-0.8	0.2	0.1	-0.8	-0.1

* statistically significant at p < 0.05; ** statistically significant at p < 0.01; *** statistically significant at p < 0.001

the control group. Positive PFP effects are seen for Manual (8.2, p<.001) and Thomas Jefferson (7.1, p<.01) on the *ITBS* Language test. The control schools also improved significantly on the Language exam, but at a slower rate (3.5, p<.001). Similar PFP effects are seen for the *ITBS* Math exam as well, with Manual student scores increasing by 4.8 NCEs per year (p<.001) more than controls and Thomas Jefferson students improving by 12.6 NCEs per year (p<.001) more than control schools.

The control schools improved significantly on all three *CSAP* exams; by 1.2 NCEs per year (p<.001) on the Reading exam, 1.7 NCEs per year (p<.001) on the Writing exam, and 1.3 NCEs per year (p<.001) on the Math exam. At Manual, pilot student scores increased faster than control scores on all three tests; the PFP effects for Manual are 1.0 (p=.3), 1.1(p=.4), and 0.5(p=.7) for Reading, Writing, and Math, respectively. These effects were not statistically different from the controls. Thomas Jefferson students experienced a statistically significant positive effect on the Reading exam of 1.5 NCEs per year (p<.05), and small and non-significant effects of 0.2 and -0.1 NCEs per year on the Writing and Math exams, respectively.

Individual Growth Modeling Analyses

There are some disadvantages to the two-stage HLM models used in this analysis. First, each year there is a different cohort of students due to movement of students: advancement from elementary to middle, middle to high school, and from grade 11 to 12, as well as transfers between schools and into and out of the Denver Public

FIG. 6-9

Unadjusted Individual Growth Models—Elementary Schools

	ITBS Reading	*ITBS* Language	*ITBS* Math	*CSAP* Reading
Control Intercept	44.5***	45.0***	44.9***	51.5***
Pilot Intercept	43.1***	45.9***	47.2***	53.5***
Difference between Pilot and Control Intercepts	-1.3	0.9	2.2	2.0
Control Slope	0.8***	0.2*	-0.5***	1.0***
Pilot Slope	0.4*	-0.9***	-1.3***	0.5**
Difference between Pilot and Control Slopes	-0.4*	-1.1***	-0.8***	-0.5*

* statistically significant at p < 0.05;
** statistically significant at p < 0.01
*** statistically significant at p < 0.001

Schools. Secondly, as seen in the adjusted elementary and high school models, student demographics and school characteristics do not fully explain differences in student achievement between schools at baseline. Individual growth modeling (IGM) is an extension of two-stage HLM in which each student is allowed to have his or her own intercept and slope. These models implicitly control for the factors outside of the school that contribute to student achievement, and in addition, smooth out some of the random year-to-year variability in each student's scores. Ideally students included in the IGM analysis should have at least three years of scores. Thomas Jefferson High School did not administer the *ITBS* Language test one year, and we have only three years of high school data in total, so Thomas Jefferson could not be included in the model for that test. In addition, the *CSAP* Writing and Math exams were not given to every grade every year, with the result that these two tests cannot be modeled using IGM at the elementary and high school levels. All IGM sample sizes are smaller due to students having incomplete testing histories or moving in or out of the study schools.

Figure 6-9 presents IGM analyses for the elementary schools, the full models are presented in *Figure A-19* of the Appendix. The elementary models confirm the negative PFP effects discussed earlier. The elementary PFP effects are estimated to be -0.4 ($p<.05$), -1.1($p<.001$), -0.8 ($p<.001$), and -0.5 ($p<.05$) for *ITBS* Reading, Language, Math, and *CSAP* Reading, respectively. The results are very similar to the HLM results, which were -0.3 ($p<.05$), -0.1($p=.7$), -0.4($p<.05$) and -0.7 ($p<.001$). The one exception is the *ITBS* Language effect which is stronger and statistically significant in the IGM analysis.

With one exception, the IGM also confirms the middle school HLM results. For *ITBS* Language

FIG. 6-10

Unadjusted Individual Growth Models—Middle Schools

	ITBS Reading	*ITBS* Language	*ITBS* Math	*CSAP* Reading	*CSAP* Writing	*CSAP* Math
Control Intercept	43.7***	49.1***	40.9***	50.9***	51.7***	52.9***
Pilot Intercept	34.7***	40.9***	35.0***	42.3***	42.0***	38.8***
Difference between Pilot and Control Intercepts	-8.9	-8.1*	-5.9	-8.5	-9.7	-14.1*
Control Slope	0.1	-2.4***	2.7***	0.4***	0.01	-1.0*
Pilot Slope	0.02	-2.5***	3.2***	0.4	0.9*	3.9***
Difference between Pilot and Control Slopes	-0.1	-0.1	0.5	-0.004	0.9	4.9***

* statistically significant at p < 0.05; ** statistically significant at p < 0.01; *** statistically significant at p < 0.001

and Math and *CSAP* Reading both methods yield PFP effects that are small and not statistically significant. For *CSAP* Math the HLM estimates a PFP effect of 1.6 (p<.001), while the IGM estimates a larger effect (4.9, p<.001). The models do not agree on the *ITBS* Reading effect. The IGM estimates a non-significant PFP effect of -0.05, while the HLM estimates a significant positive effect of 1.1 NCEs per year (p<.001).

At the high school level, IGM generally confirms the *ITBS* Reading results with PFP effects estimated at 0.8 (p=.3) for Manual High and -0.8 for Thomas Jefferson High (p=.08). HLM produced larger estimates of 1.9 (p=.09) and -1.6 (p<.05), respectively. The direction of the effects is the same for the two models, and the difference in size of the estimate is not of practical importance. For *ITBS* Language, IGM produced an estimate of 4.3 (p=.07) while HLM estimated 8.2 (p<.001). The IGM model confirms the *ITBS* Math results for Manual High. The IGM model PFP effect is 2.9 (p<.05), while the HLM effect is 4.8 (p<.001); both methods produce statistically significant results. The Manual High *CSAP* Reading effect is larger and statistically significant (2.5, p<.001) as compared to the HLM effect (1.0, p=.3).

A discrepancy occurs for Thomas Jefferson in the effect estimates for *ITBS* Math; the IGM estimates the PFP effect to be -2 (p<.05) while the HLM estimates 12.6 (p<.001). *CSAP* Reading also produced different results, the IGM estimates a PFP effect of -0.03 (p=.9) while the HLM estimates 1.5 (p<.05). It is possible that these discrepancies are due to the small number of Thomas Jefferson students that could be included in the IGM analysis.

FIG. 6-11

Unadjusted Individual Growth Models—High Schools

	ITBS Reading	*ITBS* Language	*ITBS* Math	*CSAP* Reading
Control Intercept	48.3***	44.5***	48.0***	53.4***
Manual HS Intercept	34.2***	29.2***	33.0***	30.3***
Jefferson HS Intercept	57.6***		57.9***	60.1***
Difference between Manual and Control Intercepts	-14.1*	-15.3*	-15.0*	-23.1**
Difference between Jefferson and Control Intercepts	9.2		9.9	6.7
Control Slope	0.4*	0.2	-0.2	-2.2***
Manual HS Slope	1.2	4.5	2.7*	0.3
Jefferson HS Slope	-0.5		-2.9**	-2.2**
Difference between Manual and Control Slopes	0.8	4.3	2.9*	2.5**
Difference between Jefferson and Control Slopes	-0.9		-2.7*	-0.03

* statistically significant at p < 0.05
** statistically significant at p < 0.01
*** statistically significant at p < 0.001

D. Correlation of Student Achievement with Rubric Levels

Methodology and Findings

To explore the relationship between the quality of objectives (as measured on the rubric discussed in Chapter IV) and student achievement, two-stage HLM was run on treatment period data for the pilot students, controlling for baseline achievement level. Once again three models were run for each test—unadjusted, adjusted for school characteristics, and adjusted for school and student characteristics. At the elementary school level a quadratic equation for time fit the data better than a simple linear equation. The quadratic equation allows the relationship between rubric level and student achievement to be a curved line. At the secondary level a simpler linear model was used because fewer years of data were available. The two pilot high schools are estimated in separate models. With only one school in each model, standard linear models (e.g., least squares linear model) are used, since there is no need for a two-stage model.

In the analysis files, elementary students are assigned one classroom teacher (although the student may also be taught by other teachers during the day). At the middle schools and high schools students have multiple teachers over the course of a school year. CTAC classified teachers as language arts (reading, writing, and literature courses) or math (math and computer science courses). For this analysis, one language arts teacher and one math teacher were selected randomly for each student. This approach eliminates the inappropriate use of one student's scores for multiple teachers, but it also makes it harder to detect a relationship between rubric level and achievement at the secondary level. A student taught by a teacher with a rubric level of 1 may also have other teachers for other related classes with higher rubric levels. The secondary level analyses are therefore biased against finding significant differences in student achievement between rubric levels. We explored the possibility of using the highest rubric level from all of a student's teachers; however, this produced a file with no rubric levels below 3.

Level 1:

$$\gamma ij = \alpha j + \beta 1\,(\text{Rubric } 4_{ij}) + \beta 2(\text{Rubric } 3_{ij}) + \beta 3(\text{Rubric } 2_{ij}) + \beta 4(\text{Time}_{ij}) + \beta 5(\text{Time}^2_{ij}) + \beta 6(\text{Rubric 4 x Time}_{ij}) + \beta 7(\text{Rubric 3 x Time}_{ij}) + \beta 8(\text{Rubric 2 x Time}_{ij}) + \beta 9(\text{Rubric 4 x Time}^2_{ij}) + \beta 10(\text{Rubric 3 x Time}^2_{ij}) + \beta 11(\text{Rubric 2 x Time}^2_{ij}) + \beta 12(\text{Last Score}) + \beta 13(\text{SES}_{ij}) + \beta 14(\text{Disabled}_{ij}) + \beta 15(\text{Retained a Grade}_{ij}) + \beta 16(\text{Not Proficient}_{ij}) + \beta 17(\text{Bilingual}_{ij}) + \beta 18(\text{Native American}_{ij}) + \beta 19(\text{Black}_{ij}) + \beta 20(\text{Asian}_{ij}) + \beta 21(\text{Hispanic}_{ij}) + \beta 22(\text{Male}_{ij}) + r_{ij}$$

where $r_{ij} \sim N(0,\sigma^2)$

The Level 2 model expresses the intercept of school *j* as the grand mean and deviations from that mean associated with school level characteristics and a random error term (ϵ_{0j}).

Level 2:

$$\alpha_j = \gamma + \beta 23(\text{Principal Years at School}_j) + \beta 24(\text{Percent Disabled}_j) + \beta 25(\text{Percent English Language Learners}_j) + \beta 26(\text{Percent Free/Reduced Lunch}_j) + \beta 27(\text{Percent Teachers not Fully Licensed}_j) + \beta 28_{04}(\text{Total Enrollment}_j) + \epsilon_j$$

where $\epsilon_j \sim N(0, \tau_{00})$

Since the slope of a quadratic equation changes over time, it is not possible to simply use the slope as a measure of PFP effect. Instead we estimate the NCE score for an average student at an average school using HLM. By testing for differences between the estimated average scores for each rubric level, we can determine whether student achievement is related to the rubric level of the students' teachers. *Figures A-20* through *A-42* present the detailed results of the rubric analysis. The results are discussed in detail in Chapter IV and shown in *Figure 4-8*.

Based on whether four traits of a quality educational objective (learning content, completeness, cohesion, and high expectations) were present, CTAC classified objectives into four levels. The highest quality objectives received a 4 (excellent), followed by 3 (acceptable), 2 (needs improvement) and 1 (lack of understanding or effort).

At the elementary level, adjusting for school and student characteristics, three of the six tests exhibit a positive and statistically significant relationship between student achievement and the teacher's highest rubric level (see *Figure 4-9* in Chapter IV). On the *ITBS* Reading test students of level 4, 3, 2, and 1 teachers averaged scores of 50.9, 49.7, 49.5, and 42.7, respectively. Students of level 4 teachers scored significantly higher (8.1, $p<.05$) than students of level 1 teachers. The differences between level 3 and level 1 (6.9, $p<.05$) and between level 2 and level 1 (6.8, $p< .05$) are also significant. Students of level 4 teachers score significantly higher than students of lower rubric levels on the *ITBS* Language and *CSAP* Math tests. On the *ITBS* Language test the level 4 average score of 56.8 was significantly higher than level 3 by 12.2 NCEs ($p<.01$), significantly higher than level 2 by 13.2 ($p<.001$), and significantly higher than level 1 by 16.9 NCEs ($p<.001$). The difference between level 4 and level 3 on the *CSAP* Math test was 3.6 ($p<.05$) and the difference between level 4 and level 2 was 4.0 ($p<.05$). *ITBS* Math shows a mixed relationship—the average NCE score is significantly higher at rubric levels 2 and 3 (47.0 and 47.1, respectively) than at rubric level 4 (39.7, $p<.001$) and level 1 (36.9, $p<.01$). Although the differences are not statistically significant, beginning at level 2 the average scores on the *CSAP* Writing test rise from 51.9,

to 52.0 and 52.4 at levels 3 and 4. The remaining test, *CSAP* Reading, shows no relationship between achievement and objective quality.

At the middle schools there are two statistically significant positive relationships (*ITBS* Math and *CSAP* Math) and one statistically significant negative relationship (*ITBS* Reading). Students of level 4 teachers on average scored 3.5 NCEs higher ($p<.05$) than students of level 2 teachers. On the *ITBS* Reading test the relationship is reversed; students of level 2 teachers scored higher on average than students of level 3 and level 4 teachers by 1.5 NCEs ($p<.05$) and 1.6 NCEs ($p=.06$), respectively. Non significant positive relationships are exhibited on the *ITBS* Language and Math and *CSAP* Reading tests. On the *CSAP* Writing test, students of level 3 teachers (44.1) outperform both level 4 (42.7, $p=.1$) and level 2 (43.8, $p=.7$).

Manual High School exhibits a positive relationship between teacher rubric levels and the student achievement level on the *ITBS* Reading test and the *CSAP* Writing test. On the *CSAP* Math test the direction of the relationship is reversed and rubric level 3 is significantly higher than rubric level 4. No significant differences exist for the remaining exams. *CSAP* Math is the only exam for which Thomas Jefferson High students show a relationship between rubric score and achievement. On that test students of rubric level 4 teachers have scores 2.7 NCEs higher ($p<.001$) than students of rubric level 3 teachers.

E. Correlation of Student Achievement with Number of Objectives Met

Two stage HLM was used to explore the relationship between student achievement and the number of objectives met.

Level 1:

$$\gamma_{ij} = \alpha_j + \beta1(\text{Met 2 Objectives}_{ij}) + \beta2(\text{Met 1 Objective}_{ij}) + \beta3(\text{Met No Objectives}_{ij}) + \beta4(\text{Time}_{ij}) + \beta5(\text{Time}^2_{ij}) + \beta6(\text{Met 2 Objectives x Time}_{ij}) + \beta7(\text{Met 1 Objective x Time}_{ij}) + \beta8(\text{Met No Objectives x Time}_{ij}) + \beta9(\text{Met 2 Objectives x Time}^2_{ij}) + \beta10(\text{Met 1 Objective x Time}^2_{ij}) + \beta11(\text{Met No Objectives x Time}^2_{ij}) + \beta12(\text{Last Score}_{ij}) + \beta13(\text{SES}_{ij}) + \beta14(\text{Disabled}_{ij}) + \beta15(\text{Retained a Grade}_{ij}) + \beta16(\text{Not Proficient}_{ij}) + \beta17(\text{Bilingual}_{ij}) + \beta18(\text{Native American}_{ij}) + \beta19(\text{Black}_{ij}) + \beta20(\text{Asian}_{ij}) + \beta21(\text{Hispanic}_{ij}) + \beta22(\text{Male}_{ij}) + r_{ij}$$

where $r_{ij} \sim N(0,\sigma^2)$

The Level 2 model expresses the intercept of school *j* as the grand mean and deviations from that mean associated with school level characteristics and a random error term (ϵ_{0j}).

Level 2:

$$\alpha_j = \gamma + \beta23(\text{Principal Years at School}_j) + \beta24(\text{Percent Disabled}_j) + \beta25(\text{Percent English Language Learners}_j) + \beta26(\text{Percent Free/Reduced Lunch}_j) + \beta27(\text{Percent Teachers not Fully Licensed}_j) + \beta28_{04}(\text{Total Enrollment}_j) + \epsilon_j$$

where $\epsilon_j \sim N(0, \tau_{00})$

As in the rubric analysis, the elementary models use a quadratic function for time while at the secondary levels time is treated as a linear function. The high school models were run separately for Manual and Thomas Jefferson High Schools, so those models employ a simple linear regression methodology. No models were run for Thomas Jefferson High School on *ITBS* Language, because testing rates were too low, or on *CSAP* Writing, because all of the teachers linked to students with *CSAP* Writing scores met both objectives.

The results are reported in full detail in the appendix, *Figures A-43* through *A-65*. The findings are discussed in Chapter IV and summarized in *Figure 4-11*.

Objectives were judged to have been met if a teacher submitted evidence that students had met the achievement goals set by the objective. Although the measures used by the objectives were not necessarily either the *ITBS* or the *CSAP*, this analysis shows that student achievement increases as the number of objectives met increases. The relationship is, however, complicated by the fact that teachers who set very challenging goals may not meet them even though they have the same positive impact on their students as another teacher who met both objectives but set less ambitious goals.

At the elementary level students of teachers who met two objectives had higher scores than

students of teachers who met only one objective, with differences of 2.1 ($p<.001$), 1.9 ($p<.01$), 3.3 ($p<.001$) on the *ITBS* Reading, Language, and Math exams (see *Figure 4-11* in Chapter IV). The same was true of the *CSAP* Reading, Writing, and Math exams with differences of 2.1 ($p<.001$), 0.5 ($p=.5$), and 3.9 ($p<.001$), respectively. In addition, on three tests (*ITBS* Language and *CSAP* Reading and Math) the average scores differed significantly between students of teachers who met two objectives and students of teachers who met no objectives. For the remaining three tests, there was no statistical differences between the students of teachers who met two objectives and students of teachers who met no objectives.

This analysis uses the same randomly selected secondary teachers as the rubric analysis, and thus the secondary results are again biased toward finding no significant differences. At the middle school level on the *ITBS* Language and Math and the *CSAP* Math tests, meeting either one or two objectives was associated with higher scores than meeting no objectives, but the difference was only statistically significant for the *ITBS* Language test. On the *ITBS* Reading test the students of teachers who met one objective had scores 1.8 NCEs lower than students of teachers who met two objectives ($p<.05$) and 2.9 NCEs lower than students of teachers who met no objectives ($p<.05$). No association between achievement and number of objectives met was detected on the *CSAP* Reading and Writing tests.

At Manual High School, achievement is higher for students of teachers who met two objectives on four tests (*ITBS* Reading and Math and *CSAP* Writing and Math). The four tests for which comparisons could be made at Thomas Jefferson High School also reveal higher achievement levels for students of teachers who met two objectives. The differences are statistically significant on the *ITBS* Reading test at Manual; the difference between meeting two objectives and meeting one and no objectives are 3.7 ($p<.05$) and 3.8 ($p<.05$), respectively. The *ITBS* Reading test also shows statistically significant results at Thomas Jefferson, with a difference of 5.2 ($p<.001$) NCEs between meeting two objectives and meeting one objective.

F. Correlation of Student Achievement with Teacher Experience and Length of Time in the Pilot

The experience levels and length of time that teachers participated in the pilot are likely to impact student achievement. To investigate whether this is true, the effects of PFP on *CSAP* Reading, *ITBS* Reading, and *ITBS* Math scores are calculated for subgroups of teachers and reported in *Figure 6-12*. The calculations are based on two stage HLM analyses which adjust for previous year's score and student characteristics. In Chapter IV Teachers-in-Residence (TIRs) were found to be more likely to write lower level objectives *(Figure 4-7)* and to be less likely to meet their objectives *(Figure 4-10)*. At the elementary level this seems to have translated into lower achievement levels for the students of pilot TIRs as compared to the students of control TIRs. On the *CSAP* Reading test the change in score over time for students of pilot TIRs was 2.5 NCEs ($p<.05$) per year lower than that of control school TIRs. The effect for the *ITBS* Math test was also negative, but not statistically significant. In contrast, the students of middle school TIRs performed better than those of control TIRs (1.5, $p<0.05$) on the *CSAP* Reading test.

Teachers with 15 or more years of experience were less likely to meet their objectives than teachers with less experience *(Figure 4-10)*. At the elementary level this translated into lower pilot student achievement compared to control students of teachers with the same level of achievement by 1 NCE per year ($p<.001$) on the *CSAP* Reading test and by 1.9 NCEs per year ($p<.001$) on the *ITBS* Math test. The effects for teachers with 11 to 14 years and more than 15 years of experience at the middle school level are positive and statistically significant for *CSAP* Reading (2.8 NCEs per year, $p<0.001$ and 1.8 NCEs per year, $p<.01$, respectively). Also at the middle school level we see that the students of the least experienced pilot teachers performed worse than the controls by 2.3 NCEs per year ($p<.01$).

The quality of objectives and the percent of objectives met improved with the length of time that teachers participated in the pilot. *Figure 6-12*

FIG. 6-12

Estimated PFP Effect by Teacher Characteristics
Weighted Two-Stage Hierarchical Linear Model Adjusted for Student Factors and Previous NCE Score

		CSAP **Reading**			***ITBS*** **Reading**			***ITBS*** **Math**		
		Effect	**P(Effect=0)**	**Number of Students**	**Effect**	**P(Effect=0)**	**Number of Students**	**Effect**	**P(Effect=0)**	**Number of Students**
Elementary School										
Teacher-in-Residence[1]		-2.5	0.0144	1039	0.9	0.3592	1303	-1.6	0.1195	1227
0-3 Years Experience[1]		0.2	0.6279	3101	0.2	0.6260	6465	-0.7	0.1449	5552
4-10 Years Experience[1]		-0.6	0.3226	1762	2.3	0.0001	2986	1.8	0.0147	2395
11-14 Years Experience[1]		0.4	0.2835	3970	-0.6	0.0929	6675	0.3	0.4654	5614
15 or more Years Experience[1]		1.0	0.0059	5601	-0.6	0.1123	8108	-1.9	0.0001	7249
Two Years Pilot Participation[2]		-0.3	0.4777		0.8	0.0177		-0.01	0.9728	
Three Years Pilot Participation[2]		-0.1	0.8005	4904	1.3	0.0010	9424	0.5	0.3125	7579
Four Years Pilot Participation[2]		0.1	0.8428		2.2	0.0002		2.7	0.0001	
Middle Schools										
Teacher in Residence[1]		1.5	0.0487	2050	-0.1	0.9259	1631	-0.2	0.9003	2103
0-3 Years Experience[1]		-2.3	0.0013	5302	-1.5	0.1188	4532	1.9	0.2695	2425
4-10 Years Experience[1]		0.1	0.9402	1908	1.2	0.4144	1692	2.9	0.1449	2211
11-14 Years Experience[1]		2.8	0.0001	5209	-0.03	0.9662	4274	2.8	0.0094	2742
15 or more Years Experience[1]		1.8	0.0019	5431	0.7	0.436	4366	-0.6	0.4231	4388
Two Years Pilot Participation[2]		2.0	0.0001	2331	-0.2	0.7140	1868	1.2	0.1084	1129
Three Years Pilot Participation[2]		3.2	0.0001		0.5	0.6337		2.8	0.0399	

[1] Effect = Difference Between Pilot Slope (Change in NCE Score over Time) and Control Slope
[2] Effect = Difference in Mean NCE Score from Mean NCE of One Year of Participation in Pilot

shows that the elementary level students of teachers with two years participation in the pilot had average scores 0.8 NCEs higher ($p<.05$) than students of teachers who had been in the pilot only one year on the *ITBS* Reading test. On the same test there was also a significant difference between three years and one year of 1.3 NCEs ($p<.001$) and between four years and one year of 2.2 NCEs ($p<.001$). On the *ITBS* Math test there was no detectable effect until the fourth year (2.7 NCEs, $p<.001$).

At the middle school level achievement scores were higher the longer students' teachers had been in the pilot on the *CSAP* Reading exam. Students of two-year teachers scored 2.0 NCEs higher ($p<.001$) on average and students of three-year teachers scored 3.2 NCEs higher ($p<.001$) than

FIG. 6-12 CONTINUED

Estimated PFP Effect by Teacher Characteristics
Weighted Two-Stage Hierarchical Linear Model Adjusted for Student Factors and Previous NCE Score

		CSAP Reading			*ITBS* Reading			*ITBS* Math		
		Effect	P(Effect=0)	Number of Students	Effect	P(Effect=0)	Number of Students	Effect	P(Effect=0)	Number of Students
High Schools										
Teacher-in-Residence[1]	Manual	2.0	0.7168	986	-4.5	0.4184	926	1.6	0.8354	1124
	Jefferson	-1.8	0.4054		-2.2	0.5107		6.8	0.0253	
0-3 Years Experience[1]	Manual	3.4	0.3083	2945	-6.8	0.2306	2745	2.8	0.7661	2014
	Jefferson	0.1	0.9752		1.7	0.3024		4.9	0.0044	
4-10 Years Experience[1]	Manual	-2.6	0.4305	1018	-6.7	0.4363	965	-13.2	0.1666	1067
	Jefferson	-6.4	0.0845		-4.0	0.3012		*	*	
11-14 Years Experience[1]	Manual	0.5	0.8850	2398	-5.9	0.1868	2478	5.2	0.1889	1513
	Jefferson	0.7	0.7228		1.1	0.6872		-0.3	0.9713	
15 or more Years Experience[1]	Manual	2.5	0.4117	4086	1.5	0.6954	3449	-1.5	0.7897	2704
	Jefferson	-1.1	0.2240		-1.3	0.2920		2.3	0.1615	
Two Years Pilot Participation[2]	Manual	1.6	0.0606	704	3.0	0.0625	702	1.8	0.1718	592
	Jefferson	0.9	0.1447	972	2.3	0.0011	1179	0.5	0.5451	857

[1] Effect = Difference Between Pilot Slope (Change in NCE Score over Time) and Control Slope
[2] Effect = Difference in Mean NCE Score from Mean NCE of One Year of Participation in Pilot
* Testing rate at Thomas Jefferson High School was too small for a reliable estimate

students of one-year teachers. The results were similar for *ITBS* Math although the two-year difference was not statistically significant. Students of two-year pilot participants scored higher on all three tests at both pilot high schools, but the difference was only statistically significant for the Thomas Jefferson *ITBS* Reading exam.

G. Summary

Adjusting for differences in school and student characteristics, the estimates of the effect of the pilot on elementary school achievement are negative and statistically significant for five of the six tests, with no effect evident on the sixth test. In interpreting these results, one must keep in mind that with the large number of observations in the sample it is very easy to detect small differences. It is also important to consider whether the differences are of practical significance. For example, the PFP effect estimate of -0.3 for the elementary *ITBS* Reading test would result in an average drop of less than 1 NCE in three years, an amount that would be judged by most researchers to be negligible.

At the middle schools we see more promising results. Both pilot and control students achieved more than a year's growth on the *CSAP* tests, with pilot students outperforming controls on the Writing and Math exams. The PFP effects are 0.7 for Writing and 1.6 for Math. The Math result, in particular, is both statistically and practically significant, since it represents an average increase of nearly 5 NCEs over a three year period.

Students at the two pilot high schools saw larger increases in *ITBS* Language and Math NCE scores than the control students. Manual High School students achieved positive but not statistically significant PFP effects for *ITBS* Reading and all three *CSAP* tests. Thomas Jefferson High students

performed significantly lower on *ITBS* Reading, but significantly higher on *CSAP* Reading than control students. Thomas Jefferson's PFP effects for *CSAP* Writing and Math are small and not statistically significant. We cannot rule out the possibility that the achievement gains seen at Manual are due at least in part to the reorganization that the school underwent simultaneously with joining the PFP pilot, however, since the results are supported by those of Thomas Jefferson High School, PFP may also have contributed to Manual's positive results.

There is convincing evidence that the highest quality rubric level (4) is correlated with higher achievement. Eight tests (three at the elementary level, two at the middle school level, and three at the high school level) exhibit a statistically significant positive difference in average achievement scores between rubric level 4 and lower rubric levels. One of these tests also showed that rubric levels 2 and 3 are statistically higher than rubric level 1. In three tests, the middle levels of the rubric are statistically higher than rubric level 4, but in two of these tests the middle rubric levels are also statistically higher than rubric level 1. In addition, five tests showed positive correlations between rubric level and achievement that are not statistically significant.

There is also evidence that having a teacher who met two objectives is associated with higher average NCE scores at the elementary, middle and high schools. As is the case with the rubric analysis, the strongest evidence comes from the elementary schools, where the effects are not diluted by multiple teachers per student.

Teachers-in-Residence were found to write lower quality objectives, and to be less likely to meet those objectives. Comparing pilot TIRs to control TIRs, we find that the PFP effect is negative at the elementary level but positive at the middle school level. Similarly, we see that the PFP effect for teachers with over 15 years of experience is negative at the elementary level but positive at the middle school level. The elementary school results reinforce the need for better objective setting support for TIRs and other less experienced teachers. At the secondary level where students are exposed to a number of teachers, more experienced teachers may compensate for any negative effect of TIRs.

Student achievement rises as length of teacher participation in the pilot rises. The increase in objective quality and percent of objectives met are being matched by increases in student achievement. This is a promising result and suggests that a sustained focus on objective setting will over time lead to improved student achievement.

CHAPTER VII

Catalyst for Change

A. Introduction

The Denver pilot has evolved in ways that have consistently tried to understand, support and reward the contributions of quality teaching to student learning. By using the progress of students as both the driver and end result, the pilot has served as a catalyst for systemic change. During the past thirty years, there has been significant national interest in school reform initiatives. Few of these have achieved the degree of reach into the system as has Pay for Performance in Denver.

The pilot's emphasis represents a departure from many earlier attempts by districts in the United States and the United Kingdom to implement some form of performance-based compensation. Their underpinning premises often derailed these efforts. Some were based on the belief that compensation is the sole or primary incentive for teachers to perform at high levels. Others were designed to be punitive, punishing teachers who were labeled as underperforming. Virtually all have been predicated on the idea that merit pay or its equivalent could be implemented without making major changes in how the school district functions. These operating premises have generally proven to be faulty.

The focus on student achievement and a teacher's contribution to such achievement can be a major trigger for change—*if the initiative also addresses the district factors that shape the schools.* For example, if the priority is on student achievement, then the district will need to develop the ability to provide schools with baseline data on student, classroom and program performance. If teachers and principals are to examine performance in the classrooms, then the district will need to provide appropriate assessments and data that follow individual student growth. If teacher contributions to student progress are to be rewarded, then the district will need the capacity to integrate the human resources and student achievement data systems. If the needs of students, teachers, principals and parents are to shape the district agenda, then the district will have to reconfigure budgetary allocations, curricular and instructional support, and professional development services.

Denver introduced Pay for Performance as a new element in a large urban setting. The pilot has been a catalyst for changing the district so that it could become focused on student achievement in a more coordinated and consolidated way as required by Pay for Performance. A key part of Denver's story is how a pilot, a subsystem functioning with a sense of urgency, engendered positive change in a larger institution. Many of the changes have been systemic—changing how the system thinks and behaves. They remain, though, works in progress.

This chapter highlights areas of change which have been significant and often subtle. The chapter also identifies gaps and circumstances which are part of the challenge which lies ahead for the district.

B. The Board of Education and the Denver Classroom Teachers Association

Under Pay for Performance, an unusual form of partnership emerged between the Board of Education and the Denver Classroom Teachers Association. There are numerous examples of collaboration between boards of education and unions on issues related to power sharing. Indeed, through much of the 1990s, many districts throughout the United States engaged in various forms of shared decision-making. Differing markedly from these efforts, Denver leaders came together to collaborate and take risks on behalf of student achievement.

The Denver Board of Education has been steadfast in supporting the implementation of the pilot. Some board members initially had a level of concern about the union's commitment to the pilot. However, over the course of the pilot, all board members came to value the new level and form of collaboration which comprised Pay for Performance. One board member notes, "It has been a forum for the district and the union to work collaboratively and develop trust in one another. This is a very different relationship from the past, and Pay for Performance has given us that opportunity." Another board member adds, "The Board of Education and the Association are now working together on PFP; this is a lesson.... It's working together on behalf of kids.... It will help us move closer to student achievement." When discussing the impact of the pilot, yet another board member stresses, "Strengthening the relationship [with the union] is the most important thing." The view of the board overall is summarized by one member, "The first lesson we learned is that you can partner with teachers. It is possible to change the way business is done."

The Denver Classroom Teachers Association has also made a serious commitment to this collaboration. A key leader comments, "DCTA has placed this project as our top priority. We have placed our best people and given most of their time to the successful completion of the pilot." A leader notes, "This project has been the single best effort the Denver Public Schools and the union have been involved in—without exception." Another leader adds, "The collaboration has been amazing."

One of the collaboration's most pivotal trials came early in the life of Pay for Performance. From the time the pilot was formulated in early 1999 through Spring 2001, the superintendency of the Denver Public Schools changed five times. This turnover is described in detail in *Pathway to Results.* Having five chief executive officers in a two-year period would derail most organizations and, certainly, most new initiatives. Many school reforms have been undercut by far less dramatic events. In contrast, the Denver Board of Education and the Association worked together. They not only ensured that Pay for Performance would remain a priority during this period of leadership turmoil, but helped the pilot to achieve even greater organizational reach.

This board/union collaboration on behalf of student achievement is one of the significant developments resulting from the pilot. It is also one of the most tentative. If the collaboration is not nurtured carefully and extended to other parts of district conduct and operations, it can easily fall victim to the divisiveness among boards, unions and districts that characterize much of urban education. The pilot has demonstrated a better way of conducting business.

C. Focus on Student Achievement

The pilot has significantly increased the school and district focus on student achievement. This focus has grown with each succeeding year of pilot implementation. It is a trend that is identifiable in both survey responses and interviews.

More than 70% of survey respondents have consistently indicated that student achievement is a goal of the pilot. Perceptions of the *increased focus* on student achievement, however, have changed over the course of the pilot. In 2001, 47.5% of respondents agreed that "Pay for Performance had led to a greater focus on student achievement at my school," while 52.5% disagreed. In 2002, respondents were asked if they thought their school's focus on student achievement had changed. In response, 57.4% indicated that it had improved, while 40.1% noted that it had stayed the same. In 2003, when asked about the impact of Pay for Performance, 68.5% of the respondents indicated that the pilot had had a positive impact on their "school's focus on student achievement" compared to 29.9% who felt that the pilot had not affected this focus. The 2003 responses are particularly noteworthy when viewed by school; all but one of the pilot schools believed that Pay for Performance had a positive impact on the respective school's focus on student achievement.

As Pay for Performance became more familiar to pilot participants, and as the implementation of the pilot was continuously strengthened, the focus on student achievement became a growing reality at the pilot schools and affected practices as well as perceptions. A pilot teacher says, "I think [PFP] has brought teachers and administrators together working on the learning process."

Through the pilot, an emphasis on understanding individual student growth emerged at the school sites. This was a necessity for implementing Pay for Performance at the pilot schools. This emphasis subsequently expanded to other schools in the district. It was later reinforced by other district initiatives and by the requirements of the No Child Left Behind Act.

The focus on individual student growth had implications for the classrooms. As described in Chapter V, teachers and principals were better able to set objectives that were based on the learning needs of students. At many of the pilot schools, the objective setting went from being an initial exercise in writing to more of a practice of thinking differently about instruction. In addition, there were changes in how teachers approached the meeting of these learning needs. One pilot teacher indicates: "There are many positives with PFP in our school. School-wide focus is one of them. I feel this had a positive effect on our school." Another pilot teacher adds: "It helps focus on achievement through the year. It helps teachers plan towards a goal in increments." As they were able to examine and understand their students' progress differently, it was easier for teachers to focus more on meeting the needs of individual students.

At the control schools, there were principals and teachers who began using the pilot's template for preparing objectives as a tool to help them to focus on student achievement. They indicated that they were going ahead with this practice even before it was instituted by the district. In such ways, the pilot's emphasis on student achievement has seeped into the system and has created more demand for tools that would help schools to act on this priority.

While there are varying opinions in the central administration as to the advisability of adopting Pay for Performance into the district, there are many in the administration who feel the pilot has moved the district toward a clearer focus. "There has been a greater focus on student achievement because of the pilot," a key leader notes. Another adds, "Pay for Performance has started us thinking outside the box.... It has helped us understand why we must focus on accurate data concerning teachers and students. It has shown us how important the setting of objectives is to improving student achievement. PFP has helped us build energy around accountability and student achievement."

This focus on student achievement has progressed and been reinforced over the four years of the pilot. It represents a serious step forward for the district.

D. Shaping Implementation: The Role of Teachers and Principals

Teachers and principals were provided with multiple opportunities to influence the course of the pilot. For many, this was a marked departure from past district practice. As in other large districts, Denver's site level practitioners characteristically described past reform and improvement efforts as being done to them, rather than with them. Others described what they perceived as a repeated pattern of the needs and priorities of the sites being

overlooked by district initiatives. In contrast, the Pay for Performance pilot and study made it possible for the voices of teachers and principals to be heard and acknowledged.

Due to the construct of the pilot, the activism of the Design Team and the application of research findings and related technical assistance, teachers and principals were able to become active shapers, instead of being passive beneficiaries or victims, of the pilot. This involvement began at the very start and continued throughout the four years of the pilot.

Teachers made clear through their comments that their involvement would be active. As detailed in Chapter II, the participation of the pilot schools was based on faculty votes. Teachers indicated they wanted "a chance to have a say," "to be part of the pilot for input," and "to prove if PFP can or cannot be fairly implemented." One teacher indicated, "This process is the future. We might as well do it first." Another added, "I wanted to participate in figuring out how to make it work." At one site, a teacher commented, "I wanted to be in on the design of the project—to be able to have input rather than be told how it will be a few years down the road." One teacher mentioned, "The fact that we are an 'at risk' school—I wanted a real world account of the PFP program to be put into the records." Another stated, "The staff is a confident bunch and figured, if this would be implemented in the future, we should have a hand in shaping it."

The practitioners at the pilot schools used the full range of pilot-provided vehicles to make their voices heard. In particular, these included the study's interviews, surveys and classroom observations, regular sessions and discussions with the Design Team, and foundation-sponsored events.

Many teachers and principals thought carefully about how to shape the pilot. A teacher stated, "Just being able to be heard has been so important to my teaching and practice." Initially, many teachers and principals just wanted to master the basic mechanics of the pilot. However, as the pilot evolved, teachers and principals grew in their understanding of what they needed to be successful in their classrooms and schools. Increasingly, as true shapers, they identified gaps in the system and needs that had to be addressed. They began placing demands for supports that would enhance their work with children. Comments ranged from "I need an assessment that complements *CSAP*," to "the curriculum needs to be aligned with the assessments," from "there have to be multiple measures to be fair," to "I need to know what a good percentage for gain would be."

Site level practitioners made these needs visible during the pilot's four years. They are not isolated issues; school staff experience similar needs throughout the district. The pilot and study provided the vehicle for the concerns of the sites to begin to drive district actions. This helped both the school sites and the central administration to become more sensitive to the needs of the classrooms.

E. Third Parties

The national track record on reform shows that the participation of external parties can be helpful to school and district improvement efforts. It has proven most pivotal when focused on issues of real import to students, practitioners and communities. It has proven far less helpful when it supports piecemeal programs.

When the pilot was created, the sponsoring parties agreed that they would seek external funding partners, technical assistance and research support. They knew that Pay for Performance was a high stakes undertaking for the district and wanted to maximize the results that would come from the pilot. Although there can be a tendency for any large organization—particularly a public institution—to be somewhat xenophobic, it is not unusual for a school district to seek outside financial and technical support. Using external research to help a district be open to an honest mirror is unusual. The Board of Education and the Association felt that it would be important to the pilot. Later, due to the commitment of the Design Team, it became part of a larger district interest in becoming more of a learning organization.

There is an inside-outside dimension to effective school reform. Simply put, educational reform has proven extremely difficult to achieve without outside help. There are two essential reasons for this circumstance. First, due to the entrenched nature of large bureaucracies, internal reformers need to be bulwarked by external

advocates. Second, due to the complexity of the issues affecting public education, a broader range of expertise is needed than can be found solely in districts.

As detailed in Chapter X, foundations took significant risks in supporting the field testing and study of an unproven venture in linking student achievement, in part, to teacher compensation. As key participants and partners, the foundations proved one of the most critical third parties.

The impact of the third parties was "really positive in the end. Everybody has a different push on the district. Without the third party support, the union and district couldn't have done it," notes a leading philanthropic supporter. Another foundation leader adds:

> "I don't think we would have the pilot today if DPS didn't think they were being watched. This project cannot drift into the night because people are watching what is happening. Having CTAC and the foundations involved brings in a significant level of accountability. They provide accountability. The district can't drift quietly into the night. The third parties know their stuff."

While the third parties functioned as the conscience of the pilot, they were also willing to venture into new educational terrain. Another foundation leader comments, "This project calls for taking huge risks on everyone's part. Anything I can do to help the project, I am willing and want to do. We will want to monitor the project as it continues. We think it is important to have a third party like CTAC involved." Yet another foundation leader states, "The impact of this project is important and designates a new time and age."

An important corporate leader feels that third parties, organizational stability and Pay for Performance have the potential to shake up the pattern of business-as-usual practices. He comments, "CTAC plays a central role in really making a difference. The biggest problem is stability with the superintendent and board. The history of educational reform is that it has been stillborn time and again and teachers are jaundiced from reform efforts." He believes that the blend of new directions, more stable leadership and third party support are encouraging to those who seek change in the district.

The third parties operated in a highly collaborative manner with the Design Team, the Board of Education through its liaison to the pilot, the administration and Association leaders. The specific entry points for the third parties varied (e.g., funder, technical assistance and research provider, communications specialist, corporate leader). However, there was common ground in their collective emphases to help the district build capacity, learn from the research study and make change. Much of this input was regularly channeled through the Design Team.

Particularly because of the involvement of the third parties, issues that long affected the district were now put on center stage. This provided the protection needed for the Board, the administration, the Association and, most pivotally, the Design Team to take action on findings and recommendations, an essential function. Due to the highly visible nature of public education, there is characteristically a de facto tendency to manage for impression rather than results. There is generally a worriment about how issues will play out in the press or the political arena. This concern leads to defensive leadership. Through the pilot, issues that needed attention began to get attention.

The district was particularly interested in learning from the pilot research findings. A series of detailed management letters were prepared by CTAC for the superintendent and circulated to the Design Team, the Board of Education and the Association. These management letters delineated emerging issues, concerns related to district capacity, and recommendations. Many of these issues and concerns are discussed in Chapter VIII. The management letters became the basis for analysis, discussion and follow-up by district and pilot leaders as well as by external funders. This same approach was used to migrate the findings and recommendations in *Pathway to Results,* the midpoint report. By taking these steps, Denver leaders were moving the system in a new direction—becoming a more research-driven district.

F. Teacher Compensation and Pay for Performance

Pay for Performance has been the catalyst for developing a fundamentally new compensation plan for teachers in Denver. This plan is nearing the final stages of development. The members of the Association and the Board of Education will vote on the plan in 2004.

In June 2000, the sponsoring parties faced a critical junction. The pilot had been embedded in the contractual agreement between the district and the Association. The final agreements in the contract had resulted from intense, eleventh hour round-the-clock negotiations. The final contractual language described the framework for Pay for Performance. It did not, though, focus on the desire or intent to develop a new compensation plan.

As described in Chapter II, the need emerged to clarify the purpose of the pilot. At a June 2000 board retreat with CTAC, board members indicated they had intended for the pilot to lead to the development of a new compensation system. Yet the pilot was built around short-term bonuses, and a new compensation system would require a special developmental effort. The ensuing discussion underscored the importance of clarifying the purpose of the pilot and addressing the issue of the development of a compensation plan. This subsequently became the joint emphasis for both the board and the union. As described further in Chapter VIII, the Joint Task Force on Teacher Compensation became the structural embodiment of this collaboration and priority.

Pursuing a new direction in a compensation system is a major undertaking in any district or community. When this involves potentially linking part of a teacher's compensation to student achievement, it is particularly significant. In Denver, it became an opportunity for multiple parties—the central administration, the union and the Design Team—to engage diverse publics to present their views on the issue of compensation. The pilot, in effect, provided the basis for engaging constituencies around a potential shift in public policy.

Teachers and administrators have used surveys, interviews and meetings with the Design Team, the Association and the Joint Task Force as ways to make their concerns known. They have indicated both their preferences and their perceptions of prevailing educational and political realities. This was exemplified in the Spring 2002 survey results of pilot schools. Sixty-one percent of the responding teachers agreed or strongly agreed that "student achievement will eventually be connected to teacher compensation in this district," while 39% disagreed or strongly disagreed. In addition, 56.1% of the teachers felt that "a teacher's contribution to student achievement should be rewarded in financial terms," whereas 43.9% disagreed or strongly disagreed. However, less than half (47%) agreed or strongly agreed that "a compensation plan that includes student achievement *could* work in this district; 53% disagreed or disagreed strongly.

The concerns of teachers and administrators about a larger scale implementation of Pay for Performance provide a roadmap of issues that the district will need to address. Their concerns cover a range of critical topics.

On the overall compensation plan, a central administrator indicates:

> "[It would work] if teachers were held accountable for 85% of their kids, were well trained and well supported . . . if the agreement were child-centered, not adult-centered, if assessments were at their fingertips, if professional development were strong—both central and site directed—and if there were parent involvement, then maybe it would work. We would need: consistency of curricular support, lots of administrative training and methodology aligned to support teachers. It all needs to be aligned with the formula for highly impacted schools."

On the link between Pay for Performance, instruction and evaluation, a pilot teacher states:

> "In order for PFP to work—for setting up goals and how to meet those goals, there are some pieces that are missing. Teachers need to be evaluated on a different level. It has to be done by people that are practitioners that know how to evaluate instruction. That piece is not there."

On the curriculum, a pilot teacher notes:

> "I have a hard time feeling that the curriculum can be standardized across the district for every

sschool in order for the work of some teachers and not others. If the district is allowed to move ahead on PFP there must be much more standardization in the instructional areas."

On the importance of examining individual student growth, a pilot teacher comments:

"Student achievement has to be looked at on an individual student basis in order to see the growth of each student. The teacher's professional development needs to be taken into account."

On an issue of fairness, a control high school teacher asks:

"Some teachers have 12 kids in class, some have 35. There is this huge question of fairness. It's not like elementary school. How can you know 160 kids?"

On the issue of mobility, a speech and language specialist questions:

"It's really a great idea to supplement our pay, but how do you measure success when a large percentage of our schools' populations moves from school to school in the course of a year."

G. Parents

Similar to many large districts, Denver's track record in parent involvement is inconsistent across sites. This gap affected parental knowledge about the pilot. One active parent confirmed the sentiments of many interviewed parents when she repeated, "I can't believe I didn't know that [it] was a pilot school. I can't believe I didn't know that my school is a pilot school."

Despite ongoing communications challenges, there were regular efforts to try to reach out more effectively to parents. In particular, parents weighed in on a possible new compensation plan. For example, in Spring 2002 and Spring 2003, they provided responses to a range of concerns.

Parents clearly identified the importance of a link between student achievement and teacher compensation. For example, 82.1% of the parent respondents agreed or strongly agreed that "a teacher's contribution to student achievement should be rewarded in financial terms," while 18.8% disagreed or strongly disagreed. Regarding whether a compensation system that includes student achievement results *could* work in this district, 77.5% agreed or strongly agreed. In addition, 70% agreed or strongly agreed that "student achievement

FIG. 7-1

Potential Effects of a New Compensation Plan Based in Part on Student Achievement

	Parents 2002			Parents 2003		
Where a new compensation plan based in part on student achievement could lead:	**Strongly Agree/ Agree**	**Strongly Disagree/ Disagree**	**Rank**	**Strongly Agree/ Agree**	**Strongly Disagree/ Disagree**	**Rank**
Improved student achievement	78.2%	21.8%	4	71.0%	29.0%	5
A greater school focus on student learning	86.9%	13.1%	1	82.5%	17.5%	2
Teachers working harder	86.9%	13.1%	2	81.8%	18.2%	3
Students working harder	64.6%	35.4%	5	66.4%	33.6%	7
Greater stress for teachers	64.4%	35.6%	6	72.3%	27.7%	4
Greater stress for students	41.8%	59.2%	8	54.0%	46.0%	8
Teaching to the test	78.4%	21.6%	3	83.2%	16.8%	1
Less attention paid to subjects not tested	59.0%	41.0%	7	70.4%	29.6%	6

should be connected to teacher compensation in this district." Further, approximately two-thirds (68.4%) of the parent respondents believed that "student achievement will eventually be connected to teacher compensation in this district."

Parents also indicated what they felt could result from a new compensation plan. As *Figure 7-1* indicates, the parental responses were consistent in sequential years.

Parents in both 2002 and 2003 felt that strongly that a greater school focus on student learning could be the result of a new compensation plan that was based, in part, on student achievement. They also felt strongly that teachers would be working harder and that they would teach to the test.

Parents also indicated what they want to see in any compensation plan. They want student achievement to be part, but not the entirety, of the plan. They want teachers to be both rewarded and held accountable. Parents are also in agreement with teachers on the use of multiple measures for compensation purposes. As an example, 94.1% of the parents and 93% of the teachers agreed or strongly agreed that there should be "more than one measure of student achievement used to determine performance."

While engaging more parents remains a work in progress in Denver, Pay for Performance provided a way for eliciting parental concerns on a potentially major new direction for union/management agreements and policies.

H. Summary

The Denver pilot has been a catalyst for change. It has led the district to pursue new directions in both process and substance. Pay for Performance has been based on an unusual leadership collaboration involving the board and union. It has brought in an array of external parties as financial supporters, advocates, technical assistance and research providers and, above all, as tough, honest mirrors.

The pilot has generated an increased focus on student achievement. In so doing, Pay for Performance has enabled the voices of practitioners to shape and influence practices and procedures. This level of change has, in turn, extended to the initiation, discussion and planning of a potential new compensation system. Whether ultimately approved or not, the process of engaging multiple publics is now a part of the Denver educational landscape.

Pay for Performance has enabled issues which have adversely affected district progress, sometimes for many years, to be put on center stage. This has engendered discussion and, frequently, action. Taking this course has helped the district to develop an increased capacity to make mid-course corrections—a rare occurrence in large school districts and a result that is difficult to achieve.

CHAPTER VIII

Organizational Alignment and System Quality

A. Introduction

A major initiative that focuses on improving student achievement—while concurrently exploring changes in the teacher compensation system—goes to the heart of the district mission and structure. In this context, a district can not achieve greater than usual results while using business-as-usual practices. Central departments, in particular, need to move beyond responding to requests and become active in reshaping their services to address the issues and impediments related to the pilot implementation at the schools. This is not an easy course of action. All departments struggle with many pressures and deadlines. Additional tasks are not always welcomed. In Denver, some central administrators had serious reservations about the viability of the pilot. Yet the creation of a board and union priority requires that these issues be resolved and that departmental priorities be reset. Otherwise, the priority has little meaning. Addressing such a priority involves taking on the serious challenge of aligning the organization, from the top-down and the bottom-up, in support of Pay for Performance or any other systemic initiative.

Educational reforms rarely challenge the core organized capacity of a school district, but more typically target programs, teacher development or, perhaps, school governance. This dimension of educational reform is both philosophical and practical in its rationale. A philosophical rationale is that a reform aimed at changing a particular aspect of student learning or feature of an educational program should begin at the school or teacher level. Practically, it is often far

easier to secure internal approval and external funding when a project is small and discrete. If there is a concern, for example, about the weakness of mathematical problem-solving in a group of students, reformers may initiate a professional development program that helps teachers to structure problem-solving lessons for students. This is a bottom-up or grass roots approach that more readily engages the primary stakeholders, teachers and students.

However, this approach to improving problem-solving in mathematics may also occur for reasons related to the larger organization. The advocates of the problem-solving reform may know that the math curriculum is not articulated or aligned with materials and assessments or that the teachers of the students lack adequate preparation in mathematical content. Yet the task of reforming these elements within the school district or state bureaucracy is daunting and therefore avoided. The result is the educational reform phenomenon often labeled as "tinkering around the edges" of the district. As a consequence, many reforms that are well-intentioned and implemented thoughtfully fall significantly short of their potential. They do not resolve or address root problems because the districts are not changed. There can never be enough fixes at the school level for a problem that originates in—and is perpetuated by—the larger organization. At best, a quality school change will capture the imagination of policymakers for a short period of time. At worst, it will die due to a lack of sustainability and institutional support.

The reverse of this scenario is the well-known horror script of top-down reforms. These are characteristically driven by both educational and political forces. At best, they are inadequately funded and ineffectively staffed. At worst, they engender minimal commitment from school staff and can derail positive improvement efforts that are already underway at the schools. Reforms that target the organization of school districts increasingly originate outside of the system—legislation, charters, vouchers, etc. Legislated reforms such as standards and assessments or key elements of No Child Left Behind can be well-founded. However, in the implementation process, they often become compliance-focused at the district level, rarely resulting in classroom improvements. How many standards-based school districts can say that all of their classrooms are standards-based? Ideally, school districts should function on behalf of the children, self-correcting and readjusting systems as needed in order to improve client services. In reality, though, large districts often function on behalf of the bureaucracy, such that change is cumbersome, ineffectual, and politically charged.

Denver Public Schools is a large district. The pilot involved a cross-section of the district's client and service base, but a small percentage (13%) of Denver's schools. Yet these schools presented, in microcosm, the challenges of the broader district. In so doing, the necessary interfaces of the pilot with the curriculum, assessment, student data, human resources and other parts of the system were complex, extensive and unexpectedly difficult. This led one central administrator to refer to the pilot as "a virus." Additionally, there was not clear direction to central administrators about the priority of Pay for Performance in the context of other district priorities.

Both central and site professionals describe the ensuing problems. "At the beginning [of PFP], there was a scramble of last minute negotiations. PFP was conceived in a rush, produced in a rush, and the labor was a mess," notes a central administrator. Another member of the administration adds, "A barrier at the beginning, the district was not sure what we were doing with [PFP]." A third central administrator comments, "With all of the different superintendents, the pilot got lurched around." The sites also saw these issues. "There was stumbling and miscommunication in the first year. We were unclear about where the process was going," says a pilot school principal. A teacher leader indicates, "We have refined the process over time. I wish what we have now was what we started with. There was a lot of trial and error along the way."

In Denver, district support systems were seriously challenged by the implementation of Pay for Performance, resulting in tensions between the pilot and the broader district. Many of these tensions were creative. Using the site visits, the recommendations in *Pathway to Results,* and the ongoing management letters as a springboard, many opportunities for change were identified and district action resulted. Where district departments responded to pilot needs, such as the development of an intranet

system, all students and teachers benefited. However, the pilot also provided opportunities to improve district systems, in the interest of all students, that were under-utilized. These form the challenges of organizational alignment which lie ahead for the district.

Many necessary interfaces were worked out over the course of the pilot, but others are farther from being resolved. This chapter examines several areas of pilot impact on key district systems and the impact of the district's response on the quality and outcomes of the pilot.

B. Leadership

Changes, Commitment and Trust

"The pilot has been the conscience of the district," notes a teacher leader. "It has also revealed the problems in the district." As described in Chapter VII, Pay for Performance has placed many key issues on center stage. It has also provided a protected arena in which to address these issues. In so doing, the pilot created many opportunities for leaders to, in fact, lead.

A series of leadership changes occurred throughout the life of the pilot. As discussed earlier, there were five superintendents or interim superintendents during the first two years of the four-year pilot. Additionally, over the pilot's full duration, there has been 65% turnover of principals in the pilot schools. There was a restructuring of the district into four areas supervised by four assistant superintendents. There was a restructuring of a large pilot high school into three small schools with three principals. Further, there were changes in senior management positions. These changes included establishing the new position of the chief academic officer, which was unfilled for a period of time.

Turnover in leadership positions, with a resultant amount of destabilizing, is a recurring problem in urban school districts. However, the number, types and levels of the changes during the life of the Denver pilot greatly complicated the implementation of Pay for Performance. It also exacerbated trust issues between pilot participants and the district because communication from the district about the priority of the pilot was inconsistent.

The representative opinions of participants and other stakeholders near the end of the pilot's third year (Spring 2002) illustrate the prevailing concerns. "PFP will not go forward in this district. Teachers don't trust the district. PFP is not a bad idea but we just don't trust the powers that be," states a pilot teacher. A pilot principal adds, "There is a confused district mission and poor communication [between administration and schools]. Now I ignore 900 Grant in order to stay focused on what is needed here."

There is a measure of tension in the relationship between any central body and its satellites. This is certainly true in the relationship between a central administration and the schools. Pay for Performance enabled these issues to move from the subterranean level of discussion to a more visible forum, requiring action. For example, a teacher leader states, "We need better leadership. Each department has an agenda but no one is in charge." Another indicates, "In the past it was the hierarchy of district structure. Now it is middle management.... They act like something awful is going to happen.... They have not taken time for the [PFP] training."

Many of the participants felt it essential for Pay for Performance to be supported. "I just hope [district leadership] gives [PFP] a chance. So many programs don't stay around long enough to know if they work or not," says a parent. An external supporter comments, "The central administration enjoys only a modicum of trust from the trenches. I'm not sure [the new superintendent] has made a dent. Five superintendents in less than three years has to be an impediment. With the changes in superintendents, people become busy thinking about other things. PFP becomes one of many important things." A teacher adds, "There has been a failure by the administration to bring the PFP concept along. This has caused a setback to be dealt with going into the election." A pilot principal feels, "There has been inconstant leadership."

The stability and commitment of quality district and school leadership during significant improvement efforts contribute markedly to the potential success of reform. On the other hand, the lack or perceived lack of commitment and support detract from potentially positive results. Remaining neutral or uninvolved is perceived as a lack of commitment. Similarly, if leadership is perceived as hedging bets—providing partial but not wholehearted support—then a mixed message is sent to the sites. Through many changes of leader-

ship, these issues shaped the landscape for the implementation of the pilot.

The perceptions about the changes in the central administration and the related operational priorities should be understood contextually. While there have been serious concerns raised about the administration, there also have been substantive contributions made by central units to the pilot. These are described throughout this chapter. The core issue regarding priorities is encapsulated by one central administrator who, while anticipating the demise of Pay for Performance, states, "In order for PFP to have worked, it should have been the primary task or focus."

The leadership of the Denver Classroom Teachers Association, the Board of Education and the Design Team has largely remained stable during the course of the pilot. The significant contributions of the union and the board are discussed in Chapter VII. In addition, the Design Team has been pivotal in developing support systems for teacher participants and building bridges between the pilot and district services. However, as noted in the mid-point report, the very presence and effectiveness of the Design Team may have led a number of central administrators to believe that they did not need to assume responsibility for the pilot.

As Pay for Performance moved forward, more central administrators did become involved in providing services and support to the pilot. There were distinct variances in the scope and extent of their involvement. For example, the current superintendent assigned several key administrators to drive major supportive changes in areas ranging from compensation planning to human resources systems. Some district administrators have worked to address specific issues, as in the loan of an assessment department staff member to work on the development of the data system that gives pilot teachers better access to student assessment data. The involvement of others, though, often has fallen short of assuming responsibility for a successful pilot implementation.

There are also issues of trust which affected the climate for implementing Pay for Performance. All school reform efforts require a measure of trust between district and school leadership and between leaders and teacher participants. Pay for Performance, which involved a significant level of risk-taking on the part of all participants, especially teachers—whose compensation was at issue—was particularly demanding of a trustworthy district leadership effort.

Issues of trust within the district have come to light at several levels. They do not just exist between the central administration and the schools. Besides the lack of trust felt by pilot participants for district leadership, interviewees have cited trust issues between teachers, and between teachers and principals. These issues are described in Chapter V. Such concerns are important particularly in light of recent research which indicates a correlation between trust and student achievement at schools.[1] The interview data near the close of the pilot indicate that little progress had been made in changing perceptions about the levels of trust in the district.

Communications

There were several dimensions to the communications strategy for Pay for Performance. These include the communications within and among the pilot schools, between the pilot and the central administration, Association and Board, and broader communication to the non-pilot schools and the community at large. These are substantive requirements of a pilot.

Different audiences had markedly different levels of understanding of the pilot. Due to their direct participation in Pay for Performance, pilot school teachers and administrators came to have the greatest understanding of the overall efforts. Control school interviews and survey data show that non-pilot teachers and principals lack fundamental information about the nature of Pay for Performance and, in some cases, are unintentionally operating on misinformation. Although buildings have union representatives and principals who share information, it has proven difficult to communicate accurately what is happening in the pilot, particularly for the level of decision making at which non-pilot teachers will have to engage. Also, as Chapter VII indicates, parents are generally not well informed about Pay for Performance, regardless of the schools their children attend.

Recent research shows that teachers who actually engage in performance pay efforts are

far less fearful of their impact and more open to changes in compensation systems than teachers who have not participated.[2] Non-pilot teachers have not reached the same level of understanding or trust of a compensation system based in part, on student achievement as have pilot participants. Further, since these schools did not elect to participate in the pilot, there is likely less openness to such a change at the outset. So the non-pilot schools, by nature, are somewhat of a difficult audience.

Interview responses by the end of the pilot show that a range of practitioners feel there have not been sufficient communications from the district in support of the pilot. A teacher leader comments, "District communications are flawed. What the principal is hearing and [what] I am hearing is different." A central administrator adds, "I don't really have enough information about PFP." The challenges of communicating a major initiative to a district and community are significant. While the pilot undertook many communications efforts, greater organizational alignment in support of these would have benefited Pay for Performance.

The district will need to continue to explore ways to communicate to all of its constituents the importance and potential of Pay for Performance to improving the education of Denver's youngsters. However, the remaining challenges should not obscure a key fact. As one control school principal discussed, "perhaps the most impressive communication about the pilot is the fact that the district, which is known for not sustaining initiatives, has stayed with the pilot for four years."

C. Structure of the Pilot Leadership

Design Team

The Design Team members were charged with implementing the pilot. Their role is described in detail in *Pathway to Results* and Chapter II of this report. Throughout the pilot, the Design Team's four members approached the work of Pay for Performance with a sense of passion, commitment and urgency. This was, alternately, both facilitated and exacerbated by the Design Team's place in the district's organizational structure.

While the Design Team began by focusing on getting the pilot started at the initial schools, its scope of responsibilities soon increased. The Design Team became the fulcrum for working with internal leaders and departments. It was also the primary point of contact for external supporters, including funders, research and technical assistance providers, and communications specialists.

Throughout the pilot, the Design Team continued to refine the objective setting and pilot support processes (see Chapter IV), promoting improvements that many participants appreciated but that some interpreted negatively. "The pilot has changed so much since inception, what its goals are. One problem is that the changes weren't clearly articulated. Few people knew what was really happening. PFP is a moving target. They are changing their minds about what it is about," a central administrator critiques. A pilot teacher adds "The DT should stop continuously making changes to the stated desires for objective setting." Another pilot teacher comments, "At the beginning we didn't get a lot of explanation and help with what we were supposed to be doing and what was expected. It has gotten better."

Working collaboratively with internal and external allies, the Design Team pushed for district changes on issues related to assessments, data capacity, professional development, and others. Over the course of the pilot, the Design Team pursued these activities in ways that did not fit neatly within the district's organizational structure. During different administrations, the Design Team reported formally or on a de facto basis to the superintendent, the pilot champions and/or the chief academic officer. It was not a traditional fit with the organizational chart of the district.

The Design Team's ability to operate flexibly—essentially outside of the organizational chart of the district—has been a double-edged sword. A central administrator comments, "There need to be tighter links to the line operative. [On the other hand] it allowed the pilot to experiment by being off line. There are pros and cons. It gave freedom by not being buried in day-to-day [line] responsibilities." The Design Team was able to advocate with many central units for greater pilot support. Yet it had little authority with district departments whose work affected the implementation and the study of the pilot. Also, there were several central administrative interviewees who stated that the Design Team, rather than their respective

departments, was funded to advance Pay for Performance.

Leadership Team

Determining effective ways to involve district and community leaders was an ongoing challenge for the pilot. There were a few false starts before an effective vehicle was identified. An initial steering committee was formed in November 2000. This structure was soon followed by a more integrated approach—the Leadership Team—in June 2001, one which directly involved key internal and external partners. As this approach was honed and focused, the broader leadership functions began to be separated from the day-to-day functions of implementation. An effort was made to create an agenda that invited key partners to provide support and react to the direction of the project. The meetings involved leaders from the Association, the Board of Education, the district, the funding community, the Design Team and the Community Training and Assistance Center (as the pilot's research arm). Even though there were changes in personnel, the different entities met on a regular basis to serve as a resource for the pilot.

Joint Task Force on Teacher Compensation

As reported in Chapter VII, members of the Board of Education indicated during the June 2000 retreat that a core intent of the pilot was to develop a new salary schedule for teachers that in part links student achievement and teacher compensation. The Association shared this interest.

Later in that same year, CTAC sent a management letter to the Board of Education, indicating:

> If this is the central purpose of the pilot, we recommend the formation of a Joint Task Force on Teacher Compensation. This is an issue area in which many districts throughout the country are experimenting. Denver should maximize the opportunity to learn from these efforts. This task force should review the national efforts at compensation systems that are based on student achievement and/or teacher performance.
>
> This task force should be discrete and separate from the collective bargaining process. It should have representation from the board, administration, teachers association, and Design Team. Its numbers should be limited, and its role should be advisory. The task force can play a critical role in sifting through the options available, identifying their strengths and weaknesses, successes and failures, and recommending possibilities to the board and the association as the pilot advances.
>
> The learnings of the task force and the findings that result from Denver's pilot can then be channeled into the collective bargaining process. This would help inform Denver's efforts to develop a new salary structure.

The Joint Task Force for Teacher Compensation was not part of the original agreement on Pay for Performance between the Board of Education and the Association. However, acting on the above recommendation, the Task Force was created by a side agreement and approved by the two parties as a companion entity to the Pay for Performance pilot. This was another example of the pilot sponsors making a necessary and strategic mid-course correction to advance Pay for Performance.

The purpose of the Task Force was to design and recommend a compensation plan for voting approval by the Board of Education and the members of the Association. The participation was generally as recommended, and community members also served as members. The Task Force subsequently became a critical component of the reform process in Denver. The vote on the new compensation plan will take place in 2004.

D. Data Capacity

Student Data Information

The data system is a pivotal component of both Pay for Performance and district management. As described in *Pathway to Results*, the pilot began without having established baseline data or a timeline sufficient for longitudinal study. Consequently, the pilot would have been unable to benchmark progress or conduct trend analyses. These were early learnings for the pilot. When technical assistance providers identified these issues, the Design Team made recommendations for change to the Board of Education and the Association. The pilot sponsors

then designated a baseline year for the pilot and extended the pilot's duration to four years.

The use of baseline data on individual student performance is a foundation of Pay for Performance. Understanding the contributions of a teacher starts with a rigorous analysis of the data on each student's individual performance. Without a reliable bank of such information, teachers are unable to set targets for student gain based on student achievement data. Particularly in a district where there is a great deal of student diversity and where many classes are heterogeneously grouped, a particular concern for teachers is the importance of having data that delineates individual student growth. For teachers, this is a fundamental issue of fairness. Indeed, in the 2003 survey responses, 93% of pilot teachers agree or strongly agree that in a compensation plan based, in part, on student achievement, "each student's growth [should be] measured from his or her starting point at the beginning of the year."

In responding to the needs of the pilot teachers and principals, the Assessment and Testing Department of the district worked with the Design Team to develop the On-Line Assessment Scores Information System (OASIS). This is an intranet system which provides assessment data on students from previous years, is customizable, and delivers scores for all students in a class to a teacher's desktop. In addition, Assessment and Testing developed a specific input system for the teacher objectives where the teacher logs on to enter their objective information in the fall. Using this system, principals can also have access to the objectives for their schools in order to finalize the objectives for the year. Also, the Assessment and Testing web site provides a user-friendly tutorial on practical applications of assessment data for teachers.

These efforts form some of the most powerful district responses to the pilot. Some, but not all, of the teachers in the pilot reported having a principal who provided beginning-of-school data in hard copies for teacher use, but most have greatly appreciated the electronic accessibility of this information for classroom planning purposes. Also, the ability of teachers to input the information on their objectives online greatly increased the accuracy of the objectives, prompting teachers to be more complete. Finally, these systems can be used by non-pilot school teachers and principals, a contribution to the entire district.

The survey responses particularly underscore the growing value of the access to and use of student achievement data. In Spring 2002, approximately half of the pilot teacher respondents reported that improvements were related to knowledge, understanding, and use of student achievement data. As examples, 51% felt that their "knowledge and understanding of student achievement data" improved; 51% believed that their "use of student achievement data to set objectives" improved; 50% indicated that "my school's use of data in setting objectives" improved; and 47% indicated improvements in "my use of student achievement data to plan instruction." In the Spring 2003 survey, 67% of the respondents saw PFP as having a positive impact on their use of student achievement data, 65% indicated a positive impact on their understanding of student achievement data, and 62% saw a positive impact in their timely access to student achievement data.

As with most innovations that schools actually use, there were recommendations for improvement. Pilot participants made these known through interviews, surveys and on-site meetings. Concerns ranged from making item analysis information available to placing the English language learner assessment data into the system. "At present, it is not possible to do an item analysis of test data. This would cost the district money, but it would be a better investment for schools," a pilot school principal remarks. Many of the suggestions and recommendations are described in Chapters V and VII.

Link of Student Achievement to Human Resources

Linking student achievement to teacher performance requires a relational database. This means that the district must be able to tie individual students to specific teachers. This necessitates having unique teacher identification numbers that are then linked electronically to students. The awareness of this need has emerged from the pilot and study; previously, it was not part of the lexicon of the district.

There is widespread agreement within the Denver Public Schools that such a system of

teacher identifiers is a requirement of Pay for Performance—or any other initiative that examines a teacher's contribution to student achievement. However, the need for the coordination of several departments make this a complicated undertaking and other priorities have competed for staff time. A temporary fix was established for pilot and control schools to cover the period of the pilot and study. However, as Pay for Performance goes to scale in the district, the importance of this gap in the district's data capacity will become more pronounced.

More than just inhibiting the expansion of Pay for Performance, this gap will prevent the district from accurately tracking the effectiveness of programs and staffing, and from meeting the reporting requirements of No Child Left Behind. It also constrains the ability to conduct high quality cost-effectiveness studies. Moreover, based on the databases made available for this study, it appears that the data about teachers—credentials, years of experience, school and class assignment, etc.—show inaccuracies and inconsistencies in different databases. The need to address these issues district-wide is paramount.

E. Quality and Alignment of Assessments

Assessment of student progress is the point of connection between student performance and teacher performance—the linkage around which Pay for Performance is constructed. Accordingly, the pilot provided opportunity for the district to approach the use of assessments more carefully and thoroughly. Numerous assessment issues have emerged.

The district has subject standards with grade level benchmarks for all subjects and a list of available assessments, but there is not an alignment between standards and assessments except for some subject areas where there are district-developed end-of-year or end-of-course assessments. This lack of alignment results in limitations on the district's ability to ascertain progress.

During the course of the pilot, all schools in the district did not administer the *Iowa Test of Basic Skills* consistently. As described in *Pathway to Results*, it was required of the pilot and control schools for this study. A related area of concern for the study and for the validity of any test administered in the district is the high number of students not assessed. This is discussed more extensively in Chapter VI. The *ITBS* is currently the district's only longitudinal student achievement database. It has recently been eliminated by the district from the testing lexicon for all schools, as has the *6+1 Trait Writing Sample*. Since the *CSAP* is only beginning to assess all grade levels in reading and writing, and only assessed mathematics in grades 5 and 8, the district is losing its capacity to follow student achievement longitudinally. This gap is a serious organizational constraint.

Within the pilot, assessment-related concerns were manifest. For example, in examining the year four teacher objectives, a total of 166 different, identifiable assessments are used to measure progress (an increase of 19% from the total of 139 in year three). This does not include 256 teachers (an increase of 60% from the total of 160 in year three) who list "teacher-made test," "criterion-referenced test," or "pre/post" as their form of measurement. As a result, the actual number of assessments used is likely to be significantly higher than the identified number. Further, the majority of assessments used have been identified as being "teacher-made" and/or teacher-scored. In interviews, teachers continue to point to an inherent unfairness of an approach which involves too broad a range of non-comparable measures. The numbers indicate this is a worsening situation—particularly when these assessments are used for compensation purposes.

Over the course of the pilot, schools have been implementing the *Colorado Student Assessment Program*. There are many concerns specific to this assessment, including the late availability of the assessment results, the use of the test to make comparisons among schools by the State of Colorado, and the stress created in the schools by the focus on improving *CSAP* scores. The *CSAP* is described further in Chapter III of this report and in *Pathway to Results.*

During each year of the pilot, teachers, principals and central administrators have described their concerns regarding district and school assessments, the administration and scoring of the assessments, and the setting of appropriately rigorous growth targets. The concerns fall into the following categories:

- the need for standard assessments for all areas and subjects before the implementation of a new compensation plan.
- the lack of assessments that are culturally appropriate for Denver children.
- the need for assessments to be administered and scored independently of the teacher.
- the need for more than one measure or multiple measures to determine student growth.
- the need for multiple years of data on students per teacher to provide greater accuracy and reliability of results.
- the lack of valid, reliable, and aligned measures for specials and specialists.
- the need for greater precision and rigor in setting growth targets.
- the lack of consistent assessments for K-2 and grade 12.

Participants have continuously suggested improvements for the use of assessments in the pilot and the broader district. "Assessments should be in line with what we are doing in the classroom," says a classroom teacher. A pilot principal states, "I'd like an objective in every major content area. I'd like to see a mandatory connection with lesson planning." Another pilot principal discusses targets, "I would like to see guidelines changed so that teachers can't set a target lower than 80%.... Why bother setting such low objectives?" A pilot school teacher expounds on this, "We need to base [PFP] not on broad schoolwide or classroom growth, although that's what the public looks for. We need to look at each child, where they began that year and how they improved during the year."

Both the assessments and their administration draw the attention of practitioners and parents. A pilot teacher comments, "We need better assessments, more standardization of assessments. We were trained on *ITBS* and that's going away. *Six-Trait Writing* is gone. *CSAP* is there, but there's no pre and post during the year. What are we going to use now? We were using *Aprenda* for ESL students and that wasn't a good tool. QRI Reading test is a problem because it is subjective based on the teacher. Different teachers get different results." A pilot teacher expresses a repeated concern, "The way tests are administered can be subjective, can depend on the teacher. Administration of tests needs to be objective." Another pilot teacher says, "Teacher-made tests is a cop-out. With standardized tests you can't control [outcomes] and it's more objective." A pilot teacher offers a critique, "If you use QRI's, there are no checks and balances." A parent reinforces the concerns of many pilot participants, "Each child has to be looked at as an individual and measured on their growth. Otherwise, don't make the teacher accountable.... There has to be some measure for individual children and that can take a lot of time. It has to be fair to the teacher."

Numerous central administrators identify key steps for district action. "The district must develop end-of-course tests," notes a central administrator, "*CSAP* is a very important measurement [because] the results of the test will be reported to the public. The district must find tests for kindergarten and other grades...." Another central administrator adds, "There needs to be much more dialogue about testing and measures that are reliable. We need conversations that can begin to describe what an appropriate test or measure would look like." Also, practitioners throughout the district believe that the services offered by specialists are not effectively measured by the assessment system currently in place." As a special educator says, "Severe-profound, emotionally disabled... those students need to be provided with alternative measures."

Throughout the pilot's four years, participants expressed the need for the district to employ multiple measures when assessing student growth. "We need to have multiple ways [of measurement]... multiple tests but standardized across the district," notes a pilot teacher. Many teachers share this perspective, particularly if student achievement becomes part of the criteria for teacher compensation. A special education teacher comments, "I would recommend that various student assessments be used to determine achievement. If it were only based on *CSAP*, I would leave the district." These concerns are also reflected by the educational research community as stated in the *Standards for Educational and Psychological Testing* (American Educational Research Associa-

tion, American Psychological Association, & National Council on Measurement in Education, 1999.) Many of the suggestions and recommendations are described in Chapters V and VII. Standard 13.7 states:

> "In educational settings, a decision or characterization that will have major impact on a student should not be made on a simple test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision (pp. 147-148)."[3]

As a representative example of teacher responses, both pilot and control school teachers showed an 88% agreement that more than one measure of student achievement should be used to gauge student achievement in any new compensation plan. Similar percentages also want to see a compensation plan where each student's growth is measured individually from his or her previous performance. As shown in *Figure 8-1,* a plurality of teachers in both pilot and control schools agree that teacher contributions to student achievement should be assessed in multi-year increments.

These findings confirm the need for the district to explore ways to use multiple measures of student achievement at the classroom level. As discussed in Chapter V, this will enhance the confidence of teachers in the fairness of any new compensation plan that has a performance-based component. It will also help broaden the understanding of the impact of the district's educational initiatives.

The most pressing need of the organization is to address the quality, alignment, comprehensiveness, and integrity of assessments being used to measure student achievement in the district. By not taking this step, the district not only places a performance pay system at peril, but also greatly reduces the quality of information upon which many student decisions are based. Effective organizations have good information about their clients.

F. Professional Development

Professional development is a critical component of successful change. In a Pay for Performance plan, it is also critical to the success of the plan itself. The expectations of Pay for Performance include that teachers and principals obtain student achievement data, analyze the results, and tailor instruction both to the curriculum provided and the students' abilities and needs. To develop and fairly measure a teacher's instructional ability, professional development is required.

The practitioners at the pilot schools draw a significant distinction between the training provided to ensure the implementation of Pay for Performance and the professional development needed to deliver and measure improved instruction. Both require a base in student achievement. They need to be aligned and mutually reinforcing.

The Design Team's support to the pilot schools has grown in sophistication, quality and responsiveness with each year of the pilot. Numerous respondents indicated that they would welcome such support expanded even further. A pilot

FIG. 8-1

Spring Survey Items, PFP, 2002

	Control Schools		Pilot Schools	
I would like to see the following elements or criteria in any new compensation plan.	**Strongly Agree/ Agree**	**Strongly Disagree/ Disagree**	**Strongly Agree/ Agree**	**Strongly Disagree/ Disagree**
More than one measure of student achievement used to gauge performance.	88%	6%	88%	7%
Each student's growth measured individually from his or her previous performance.	88%	6%	91%	5%
Teachers' contributions to student achievement assessed in multi-year increments (not annually).	52%	42%	54%	39%

teacher notes, "Training from the Design Team has been helpful. However, there needs to be training where everyone, including administrators are included."

Teachers, in particular, are forceful when describing their perceptions of their professional development experiences. "Much of the district professional development is a song and dance—preaching to teachers. I need more professional development but not just general stuff called professional development. We need to be more specific in our work," says a pilot teacher. "We had a great number of professional development sessions at our school toward the end of the year. They were not good and someone should stop these activities until they are much improved. I did not find them helpful and I think that they are a waste of time and money," comments a pilot teacher. One control teacher indicates, "We have had to take care of our own issues. The district has not been helpful." Another commented, "I'm not thrilled about staff development. I find much of the time the meetings are repetition. It's just a circle."

As indicated above, the intent of Pay for Performance and other district improvement initiatives is for teachers to use effective classroom practices to meet their objectives and improve student achievement. Many times, this requires teachers to change or improve their current practices. This, in turn, calls for providing customized support to the teachers. The Design Team has made significant progress in emphasizing the elements of a high quality objective and stressing the use of assessment data to ascertain progress. However, it is incumbent upon the district to provide teachers with the ongoing assistance needed to have a fuller impact on the classroom. This is clearly important to district leaders; significant staff and financial resources are being allocated to professional development.

During the course of the pilot and research study, teachers have described the kinds of supports that they need to make their instructional practices effective with all students. As an example, they cite the need for structured time to learn from their colleagues as a key to changing their practices. They stress that professional development needs to be based on the specific needs of their students and schools, and provided on an ongoing basis. In effect, the teachers are describing what they feel is a needed pathway. To the extent that the district's professional development is perceived by teachers as being based on the differentiated needs and the specific student achievement levels of the individual schools and classrooms, it is likely to find an increasingly more receptive audience among the teachers. This will reinforce the priority on student achievement, while helping to build more bridges between the central administration and the schools.

As discussed in Chapter V, many pilot teachers did not interpret their participation in the pilot as requiring a change in teaching practice. Instead, the pilot construct assumed to some extent that, in the pursuit of objectives and additional compensation, teachers would improve their practice. In many instances, this has occurred. Teacher interview and survey data describe changes in a variety of areas, particularly the focus on student achievement and the use of data to plan and to intervene early with underachievers. The met/not met data also show that there is a statistically significant correlation between an elementary teacher's meeting two objectives and changes in student achievement in that teacher's class.

Had instructional professional development accompanied the objective setting, the achievement findings would likely have been more extensive. Teacher interviewees have consistently pointed out over the course of the pilot that they are "teaching as hard as they can" and that "they always give their best effort." However, some teachers have also revealed, through their objective setting and survey and interview data, that they do not think that they can be successful with all students. A number of objectives are set to exclude students who have attendance issues or diagnosed learning challenges. Because there were many exclusions, the Design Team pursued this issue and sought to increase the target levels in the objectives. By the end of the pilot, the growth targets are most often set at the 75% level. Nonetheless, teachers desire additional help to address the learning needs of all students.

Early in the pilot, funding was raised from foundations to conduct a professional development audit so that the district could assess where all

professional development dollars were going, what the needs were, and how student achievement was directly impacted. This was not carried out and still remains a serious gap in district services, not only for the pilot teachers but also for all teachers in the district.

This gap is more pronounced under the new national education law. The No Child Left Behind Act, Title II, Part A, states that professional development activities will be "regularly evaluated for their impact on increased teacher effectiveness and improved student academic achievement, with the findings used to improve the quality of professional development." It continues, "Ultimately, the program's performance will be measured by changes in student achievement over time as shown through the other NCLB reporting requirements." This law places new requirements on districts in the area of professional development. It also provides Denver with an opportunity to further align the organization in support of directions identified by the pilot and needed for all of the district's major educational initiatives.

G. Principals

The quality of interaction between the building principal and each of his or her teachers is pivotal to the success of Pay for Performance. Interview data indicate that there is a wide range of behaviors around this critical interaction. A number of teachers express a lack of trust in their building administrators, describing actions that are viewed as unhelpful or even arbitrary. Descriptions of principal processes (from both the teacher and principal perspective) show that some principals are extremely thorough and assiduous in overseeing and supporting the objectives process. Concurrently, others practice a kind of benign neglect when examining teacher objectives and evidence of student performance and, particularly, in providing feedback to teachers in timely or helpful ways. While some principals were able to give mid-year feedback to teachers on objectives, others did not review them until later in the year. Where principals have been engaged and supportive, their staffs are appreciative. Teachers particularly value learning from the principals who are able to serve as instructional leaders.

Paradoxically, principals identify a lack of clarity and direction regarding their roles in teacher objective setting. This is a source of ongoing frustration. They would like to be clearer on the scope of their authority or decision-making when approving objectives or bodies of evidence. For example, a few principals express a concern over objectives with low expectations coupled with a feeling of being powerless to do anything about them. They express a need for more support from the district in the form of professional development relating to objectives and the principal's proper role in evaluating them. Principals in both the pilot and control schools have some negative feelings about their own performance evaluation process, which may spill over into their work with teachers. Exacerbating the principal trust issue is the large turnover of principals in recent years, particularly at the pilot schools. Only five of the 16 pilot schools have had the same principal throughout the pilot.

The critical interactions between principals and teachers should be strengthened—particularly as Pay for Performance goes to scale. There is a salient need for a district-sponsored program that would further build the principals' capacities in the areas of instructional supervision and data analysis. A pilot principal adds, "We need more professional development in multicultural education as our population is changing." Principals indicate that they need support in examining teacher work—from the objectives and classroom plans to classroom observations and the evidence of attainment—and ways to provide timely, helpful feedback to teachers. In this manner, the role of the principal can be clarified, and the quality and consistency necessary for any compensation system based on student achievement can be improved.

H. School Improvement Plans

The study also examined the relationship of teacher objectives to the various school improvement plans. An analysis of the 2002-2003 pilot and control school plans reveals that the schools were working from a template that contained common elements. These elements include: an introduction; three-year goals; annual goals; data statements related to important needs and barriers

to high achievement; six areas of plan strategies; an equity statement listing the ways the school is closing the learning gap; a coordination of resources page; a plan for evaluation and monitoring progress; and a sign-off from the local Collaborative Decision-Making Committee. A key goal for most of the schools was to attain a rating of at least "average" on the State Accountability Reports, but higher performing schools had the goal of maintaining a high rating on the reports.

In the 2002-2003 pilot school plans, there is no evidence that the teacher objectives are considered as part of the strategy for improving the schools. When the plans are compared to the teacher objectives in the respective schools, it is clear that teachers in most schools used the school plan as a rationale—in general terms—for their objectives. It is not clear if they were motivated to do so through a school discussion or had been prompted by the examples provided for completing the objectives. Other rationales included general references to the literacy program and the importance of the content for students. High school teachers particularly used the latter rationale. Beyond that, though, the objectives and their learning content are not included in the strategies in the school improvement plans.

A representative sample of the current control school plans showed similarities to the pilot plans, but the control school teacher objectives make fewer specific references to the plan. Interestingly, 25% of the reviewed control school plans show that staff used the PFP objectives worksheet or a modified version of it for goal setting, increasing the use of baseline data in these schools. Though this was not required by the district, it indicates a way in which the pilot's reach extended informally to non-pilot schools.

The way these components of educational service delivery should align is described by an external community leader:

> "There need to be checks and balances on these objectives. The district has to do a better job of moving from high stakes testing to focusing on other teaching services—professional development based on objectives linked to student improvement plans and district plans. The district needs to integrate and align all of this if it's going to work."

District goals, the respective school improvement plans and the teachers' classroom objectives should be carefully aligned. Each should reflect the others and reinforce a coherent agenda for improving student achievement. In this way, the district goals provide guidance for the school system, while the needs and priorities of the schools shape the district agenda. Moreover, the learning goals, standards, curricular content, instructional strategies, assessment methodologies and support systems should be readily apparent to practitioners and supervisors—the readers of the plans and the implementers of the improvement efforts. This loop needs to be tighter in Denver.

I. Relationship of PFP to Major Goals and Initiatives

The district's two highest educational goals are to increase the achievement of all students and to bridge the gap between high- and low-achieving students. The same scope and quality of organizational alignment needed to implement Pay for Performance is required for meeting these goals.

Raising bars and bridging gaps have their starting points and end results rooted in a rigorous analysis of student achievement data. It is therefore essential to have assessments that accurately and reliably measure the progress of all students towards these goals. Achieving these outcomes also requires that the teachers and principals have the appropriate data available, and that they are able to understand and interpret the data accurately, identify student needs, set appropriate learning objectives, and structure lessons accordingly. Even excellent teachers may not have all of these skills, particularly those relating to data.

A Pay for Performance system demands that a district's standards, curriculum content, instructional delivery, professional development, data capacity, assessment, supervisory and human resources be aligned. The issue of alignment cuts to the very essence of how—and to what extent—the school district is functioning in support of student learning. This applies equally to implementing Pay for Performance, undertaking the district's major literacy and mathematics initiatives, and to meeting the requirements of No Child Left Behind. Addressing the issue of

organizational alignment is pivotal to the prospects for success of all of Denver's initiatives.

J. Broader Factors

There has been an array of broader institutional and extraneous factors that have affected the climate for implementing Pay for Performance. They have made the difficult challenges of aligning the organization in support of the pilot even more daunting. These factors, and their attitudinal underpinnings, have affected perceptions and understandings of the pilot across the district. The following are a few of the salient influences which have influenced attitudes about Pay for Performance.

CSAP and the State of Colorado

As discussed in *Pathway to Results, 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. As Colorado's largest city, Denver receives significant media attention. Accordingly, the district's scores on *CSAP*—and the state's ratings of schools based on those scores—are highly visible.

For many administrators, teachers and parents, the visibility and usages of *CSAP* have resulted in an extremely high stakes testing environment. The *CSAP* also increases the level of confusion within the district. Many teachers perceive *CSAP* as the driving force in the district and the state. Consequently, they express confusion regarding the distinction between the district's goals for the *CSAP* with the goals of Pay for Performance. In addition, the pilot's focus on individual student gain differs from the public presentations of aggregate *CSAP* scores.

There are other state-level factors that affect the climate for Pay for Performance. Administrators, teachers and parents frequently cite the state report card system and several legislative initiatives as placing additional pressures on the schools and the district.

The No Child Left Behind Act

This federal law has ushered in dramatic changes for all school districts. There are now new national requirements for districts to report highly disaggregated data on student and school performance, and teacher qualifications, to the community. This provides the district with new and increased responsibilities. However, as described in this report, many of the organizational capacities needed to support Pay for Performance are equally needed to meet the requirements of No Child Left Behind.

The Economy

Difficult economic times characteristically contribute to increased levels of stress in union/management relations. In Denver, this has particular consequences for Pay for Performance. If teachers feel that there are insufficient resources to fund the pre-existing compensation system, they are more likely to question whether there will be additional resources generated to support a new compensation plan.

Administrator Pay for Performance

At the start of the pilot, the district also launched an effort to introduce Pay for Performance for administrators. The district experienced numerous problems in the conceptualization and implementation of this effort. It was not well received, particularly by the principals. Shortly after being appointed, the current superintendent analyzed the results and ended the effort. Nonetheless, there are attitudinal vestiges remaining among the administrative staff which adversely affect their perceptions of Pay for Performance for teachers.

Leadership Turnover

As discussed, there have been many changes of leadership during the period of the pilot. While the district now has stable executive leadership, there are still concerns regarding district communications and priorities as a result of turnovers of leadership.

K. A Sense of Mission

The challenge of organizational alignment is attitudinal as well as operational. It has been described by some individuals as a struggle between "the new Denver Public Schools and the old Denver Public Schools."

Like all districts, Denver has a mission statement; Pay for Performance requires more than that. It requires a sense of mission. A central administrator describes this attitudinal requisite:

"The project is a real intellectual challenge to the district right now. No one else is doing what Denver is doing. Denver is out on the cutting edge and there are many risks in doing what we are doing. The district and union need to do what is right and do it in the right way. The foundations are comprised of local opinion leaders and they are investing heavily in DPS at this time. I'm sure these leaders are not investing their funds blindly and must feel there is much to be gained from this effort. We must be willing to see this pilot through and to do whatever it takes to show the Denver community that we have done everything possible to be successful . . . there will be many benefits to this pilot whether the parties vote it up or down."

L. Summary

The Pay for Performance pilot got underway at a time when the Denver Public Schools were particularly vulnerable due to leadership changes. While a flurry of work was underway in the pilot schools, there was minimal communication from the district about the purpose of the pilot and the district's expectations. Also, a lack of communication and direction to the senior and middle managers in the district, coupled with the empowering of the Design Team to lead the implementation, contributed to a lack of accountability for the success of the pilot among staff who have a great deal to contribute. Over the course of the pilot and with new executive leadership, district systems have become more responsive to the needs of the pilot and more apt to recognize that most of the changes needed for the pilot will be critical for the entire system.

CHAPTER IX

Issues and Recommendations

A. Introduction

The results of Pay for Performance in Denver have a human face. Students and teachers are highly visible in the design, implementation and impact of the pilot. The pilot and study involved several hundred teachers and thousands of students over a four-year period, a fact that provides a constant reminder of the purpose of district and school reforms—to help teachers become as effective as they can be and to help students reach the highest learning standards. Over four years, teachers in the pilot schools have laid out key expectations for their students in their objectives and have been open to having their objectives studied, to examining their processes in interviews and surveys, and to offering critiques and suggestions to broaden the impact of the reform.

The Denver Board of Education and the Denver Classroom Teachers Association have undertaken a courageous experiment in American public education—creating a pilot based on the linkage between student achievement and teacher compensation. They have also promoted Pay for Performance as a concept so that it has become a catalyst for systemic change. They have held both the pilot and the broader district up to a rigorous external examination of results. This openness has contributed to an organizational climate, supported by third parties and internal reformers, focused on understanding and becoming accountable for student growth and creating change that benefits students. Rather than avoiding the discussion of problem areas, challenges have became more visible and amenable to analysis and solution.

Through Pay for Performance, teachers have demonstrated that they can affect the growth of individual students positively. As often stated, teacher-developed objectives are the centerpiece of the pilot. Over the four years of the pilot, there

has been a positive relationship on many independent achievement measures between student gain and excellence of teacher objectives. In working with student data and setting targets for expected gain, pilot teachers have demonstrated the importance of science to the art of teaching. As a result of the pilot, schools and the district overall have become more focused on student achievement and on the importance of understanding individual student gain.

Findings from the study indicate that a focus on student achievement and a teacher's contribution to such achievement can stimulate needed changes in district systems that support and shape the schools. Despite turnover of district leadership and structure, the reach of the pilot has extended further into the district in each subsequent year of implementation. Through the leadership of the current administration and the activism of the Design Team, the district is beginning to take the lessons from Pay for Performance to scale.

The pilot has benefited from a top-down/bottom-up approach to implementing reform. Teachers and principals have had significant opportunities to shape the implementation of Pay for Performance. Needs of pilot teachers have provided the impetus for efforts to improve the access to student data and assessment information, the linkages between student information systems and human resources systems, the quality of professional development, and other areas of district support operations. While these activities were underway in the pilot schools and the district, the pilot gave birth to a proposed new compensation system for teachers.

The organizational gains resulting from the pilot to date are significant; they are also fragile. The gains can easily erode, particularly as attention shifts to the vote on the new compensation plan. Indeed, even were there not a pending contractual vote, the national experience in reform suggests a recurring pattern: districts review the results of their pilot efforts, indicate that they have integrated the learnings into their organizations and soon move on to the next major initiative of the day. While essentially declaring victory, districts often allow the momentum and direction of the reforms to begin to languish. Over time, systems incrementally return to earlier postures and behaviors. In effect, the potential for real learning and systemic reform is undercut by the response to near-term exigencies.

In contrast, Denver has learned that including the objective setting process of Pay for Performance as a core component of the district's operations demands that standards, curriculum content, instructional delivery, professional development, assessment, supervisory and human resources be aligned. As discussed throughout this report and the mid-point report, aligning systems in support of student achievement is key to turning a good district into an outstanding one. Not only do aligned systems ensure better results from the performance pay initiative but they will support district success in the implementation of other initiatives, such as meeting the requirements of No Child Left Behind. The findings of the Pay for Performance pilot have become a true, actionable priority in Denver. The stakes are high; Denver is becoming a district in which the achievement of 70,000 or more students is in the spotlight.

The district is taking a critical turn and has an opportunity to build on the pilot. The next step is to build demonstrably upon the organizational changes made to date—changes of thinking, practice and system capacity—so that teachers and schools can maximize gains on behalf of students, and the overall district can maximize the philanthropic investment in student achievement made through the pilot to the Denver Public Schools.

The following recommendations are based on four years of scientific research, the accomplishments and findings of the pilot, and the national track record in reform. The recommendations are all geared toward building the capacity of the district to institutionalize and expand the impact of the most critical elements of the pilot. They target improvements that are sustainable, manageable and implementable at district scale. With strong practitioner input, they target the improvements that teachers and principals believe will make a difference. Further, they strengthen both the validity and fairness of the district's continuing reform efforts. The recommendations are grouped into four interactive and interrelated topics: alignment, assessment, professional development, and leadership.

B. Recommendations

Issue One: *Alignment*

OVERVIEW

As the purpose of the district's major initiatives is to increase student achievement, the organization will benefit from continuing to align its initiatives around that goal in a clear and purposeful manner. Otherwise, even increasing the district's capacities will fall short of providing integrated support to schools and classrooms. The silo effect is all too familiar within urban districts—a plethora of individual programs and activities operating independently of one another whose sum total is less than the collective potential of the initiatives. Using the learnings from the pilot, Denver can avoid a pitfall that characteristically undercuts the potential of many reforms nationwide.

RECOMMENDED ACTION

- *Bring the objective setting to scale with instructional support.* Over the course of the pilot, there has been substantial progress in improving the quality of teacher-developed objectives. Crafting objectives is the initial step in reflecting on, planning, and delivering instruction. It is not merely a writing exercise. As this effort goes to scale, it will be important to draw on these learnings from the pilot and to align the instructional supports so that teachers are assisted in improving practice based on their knowledge about student achievement in their classrooms and the specific targets in their objectives.
- *Strengthen the linkage between classroom objectives, school improvement plans and district standards and goals.* This recommendation has structural and managerial dimensions. Structurally, to the extent that the objectives, plans and goals are mutually reinforcing, the implementation of all three will be strengthened and there will be greater clarity of purpose. Managerially, ensuring this alignment will need to be a priority for principals and the area offices. All professionals within the district should be accountable for these linkages.
- *Increase the connection between student information systems and human resources systems.* Building on the district's progress in supporting the pilot schools, this recommendation focuses on establishing greater district-wide linkages among the data systems for student assessment, planning, and human resources. Particular emphasis should be placed on: (1) assigning unique teacher identification numbers to all teachers, which will follow the teachers throughout their careers in the district; and (2) structuring systems so that students are accurately linked to teachers and so that teachers, including specialists, are accurately linked to students. These linkages are critical for any effort that seeks to examine the contribution of a teacher to student achievement and the contribution of a program or school to a teacher's success. By establishing these linkages, the district will also be better positioned to address the No Child Left Behind requirement of demonstrating that the district has highly qualified teachers.
- *Project the costs of changing internal practices and requirements.* There are direct financial costs to implementing Pay for Performance systemwide. The Joint Task Force, the district and Association leaders are all actively assessing the level and source of projected costs for a new compensation system. The momentum of the pilot needs to continue under a range of financial circumstances.

  One of the major findings from the national track record in reform is that when a district moves in new directions, it needs to give up some of the pre-existing practices and requirements that consume resources and are not consistent with the district's new directions. Denver should conduct a detailed review of existing cost centers and their impact as part of the contingency financial planning to support the new educational initiatives and performance-based compensation system. The lack of sustainability is one of the reasons teachers do not trust programmatic and compensation innovations. If the resources are not aligned longitudinally to sustain the new district directions, it may be a long time before teachers may be willing to try major student achievement and compensation reforms again.

IMPACT

Denver has evidenced wisdom about this reform. Many districts function as though reform means having problems in the near term and then not having them thereafter. True reform is a more complex and recursive process. It involves identifying and addressing problems, and then moving forward to address a more sophisticated set of problems that affect district directions. Through the pilot, confronting the challenges of organizational alignment has entered the lexicon of the Denver Public Schools and become part of a district dialogue and action to shape the future.

Issue Two: *Assessment*

OVERVIEW

A portfolio and appropriate usage of high quality assessments are the marks of a district that is aligned and accountable for its students. The pilot has uncovered many inadequacies and inconsistencies in the available assessments of student progress, which are discussed earlier in this report. The district needs to develop a coordinated system of assessments that are fair, valid and can address the dual challenges of diagnosing classroom performance and making broader comparisons across grades and schools. Using student assessments for a new purpose—compensation—also requires greater assessment and data capacity, as well as a skillfully supervised and consistently administered effort at the schools so that all students have regular assessments. The district is well armed with new information from the pilot and positioned to make inroads into an area that has beset educational reform.

RECOMMENDED ACTION

- *Expand the district's assessment strategy.* There are five parts to this recommendation. They should be addressed in a concurrent, integrated manner so that they can collectively extend the reach and strengthen the application of the district's current assessment plan. In effect, the existing plan should become a component of a more comprehensive assessment strategy. An effective overall strategy will:

  a. Delineate how approved assessments align with the district's standards and curricula and identify gaps in the assessment program. Particular attention should be paid to developing adequate assessments for secondary school subjects.

  b. Anticipate how the assessments are to be used in classroom instructional planning and school improvement planning, and identify the supports to be provided to the schools. Practitioners at the schools need opportunities to provide input into the selection and use of assessments and to receive the assessment data in appropriate forms with the assistance necessary to make effective use of the data.

  c. Define clearly who is to take the tests and supervise the implementation of consistent testing practices. For purposes of monitoring progress, evaluating programs and providing accurate information to the schools and the public, it is essential that *all* students—except those legitimately exempted—are being tested.

  d. Outline the steps that will be taken, and timelines pursued, to develop or acquire assessments that are appropriate for special student populations. This is important for both educational and compensation reasons.

  e. Require the collaboration of central administrators in the curriculum, instruction and assessment areas with grade and school level representatives so that the resultant strategy will yield valuable information regarding current performance, individual student growth and longitudinal performance trends across years.

- *Define which assessments can be used for objective setting and compensation purposes.* There continues to be a need for a rational level of prescriptive direction regarding which assessments can be used as part of any new system that involves Pay for Performance. A Pay for Performance system or companion educational initiative that has too many allowable assessments will be unmanageable, will cause discord and will fail to promote valid increases in student achievement. These problems will be exacerbated when the initiatives are implemented on a large scale.

- *Make the use of multiple measures a developmental priority.* One of the major reasons pay for performance has not been implemented successfully in other districts is the lack of a single measure that satisfies the criteria of being fair, accurate and valid in measuring student learning. For four years, Denver teachers and site administrators have been raising questions about the fairness and accuracy of individual measures. These concerns can be addressed more effectively in the future if the district carefully blends the assessments that measure different areas of student knowledge or performance.

  As recommended in *Pathway to Results,* the district should charge its academic staff with developing a means to use multiple measures at the classroom level. The importance of this recommendation needs to be underscored. The charge is for the district to develop a means to *link several assessments together* to more meaningfully identify student progress and, as a consequence, teacher performance. *The linking of these assessments is what is meant by multiple measures.* Moving the district in this direction is a key developmental task.

  Multiple measures will help the district to meet a higher standard of fairness and accuracy when examining a teacher's contribution to student achievement. Further, they will enable the district to achieve a more complete understanding of that student achievement. Until the district makes strides in converting the current collection of assessments into a system of multiple measures, the district will be vulnerable when making compensation decisions on instruments that, used singly, may be questioned regarding their validity.

- *Increase the district capacity to disaggregate and analyze student achievement data.* Particularly in the era of No Child Left Behind, the district needs to build the in-house capacity to collect, disaggregate, analyze and act on data related to student achievement and school performance. This requires expanding the ability to determine actual and relative progress—school by school, classroom by classroom, student subgroup by student subgroup, and child by child—and presenting the data in different formats for, respectively, classroom, school and district use. The analyses of these data then provide the foundation for delivering classroom instruction, developing school improvement plans and managing strategically at the district level. This data capacity is a bottom line requisite for helping students and schools to succeed.

- *Convene select urban districts to analyze and take action on problems in assessment.* As a result of the Pay for Performance pilot, Denver is positioned as a national leader in undertaking innovation in the area of tying teacher compensation, in part, to student achievement. This definitionally places Denver at the center of efforts to use assessment data for multiple purposes. The ensuing challenges that Denver faces are shared in common by other districts. Denver should use its current national involvements as a springboard and convene a small number of urban districts and assessment specialists, analyze the issues of how to use assessment data to ascertain progress and make comparisons, and determine potential collective action that could be taken to guide test developers to link their efforts more directly to growing needs of urban districts.

IMPACT

The pilot has demonstrated the importance of using student achievement data to inform instruction and guide decision-making. Taking the steps listed above will be a significant help to teachers and principals who are seeking reliable means to promote individual student growth. They also will provide the district with greater means to ascertain actual progress on student achievement and craft or correct district improvement initiatives. No Child Left Behind requires districts to provide parents with extensive data on student and school performance so that they can make decisions about schools. The better the district can understand, utilize and communicate student assessment data, the more effectively it will be able to ascertain student learning progress and meet the new national requirements.

Issue Three: *Professional Development*

OVERVIEW

Virtuosity in teaching is the goal of professional development for teachers. In order for reform to occur, schools have to be places that stimulate and support teachers. Initiatives often are based on the assumption that teachers will embrace the concept of the reform and change their practices when, in fact, they may follow their prior practices in their classrooms. From both the educational research perspective and Denver's experiences in the pilot, there are profound connections between objectives based on learning content, a teacher's subject matter knowledge, specific teaching practices, and student achievement that teachers need ongoing opportunities to pursue.

RECOMMENDED ACTION

- *Establish district standards for professional development.* The district needs to determine and communicate the process and content standards that will guide the initiation, delivery and evaluation of professional development. Denver is moving in the direction of providing standards-based instruction. Establishing the standards for professional development is a natural and necessary complement to this instructional priority. They should be tied to the Colorado Teacher Standards, research about best teaching practices, the district's curriculum standards and the assessment strategy described above, and their implementation should be evaluated regularly by the site level recipients of the professional development services. This work will result in a roadmap for providing professional development services and ensuring quality control.
- *Predicate professional development on student achievement.* The priorities for professional development need to be based on continuous reviews of student achievement results by school staffs. Such a review identifies schoolwide, classroom and individual student instructional needs and instructional areas which need to be updated or improved. This, in turn, may reveal areas in which school staff or the community may need assistance in meeting these needs. By using student achievement as both the driver and end result, this emphasis for professional development is more directly consistent with the priorities for teacher objectives and the district goals.
- *Create opportunities for teachers and principals to shape professional development.* One of the key findings from the pilot was that the ability of site practitioners to influence implementation contributed to improvements in the overall effort. This kind of involvement increases the prospects of professional development to target effectively teacher needs, school priorities, and district goals. Absent such opportunities, site practitioners are more likely to perceive the district as unresponsive and lacking in understanding of their challenges. When this occurs, teachers and principals can feel disconnected from district initiatives—even when the initiatives are well conceived. Through the study, teachers and principals also made clear that they strongly valued opportunities to work with colleagues on teaching and learning issues.

IMPACT

Taking these steps will improve the quality and increase the impact of professional development services. The standards will provide a blueprint for initiating, delivering and evaluating professional development. The focus on disaggregated student achievement data will enable the objectives and instructional supports to be targeted and accountable. Lastly, directly involving site practitioners in shaping professional development services provides a valuable bottom-up complement to top-down district initiatives, enabling teachers and principals to articulate needs and support the overall district educational agenda.

Issue Four: *Leadership*

OVERVIEW

Many reforms fail for lack of sustained leadership. The Board of Education and the Association demonstrated leadership as they joined to create the Pay for Performance pilot. The Design Team has provided creative leadership in advancing the pilot through uncharted pathways. Many teachers, principals and some key district staff have made

important leadership contributions. However, pilot findings also show that many parties were not well prepared to supervise the new objective setting or support the implementation of the objectives in the classrooms and the schools. As the effort moves forward to institutionalize the critical elements of the pilot into district practice, the vision and strength of leaders throughout the district will be essential to shape and guide the reform through its next steps.

RECOMMENDED ACTION

- *Broaden the collaboration on behalf of student achievement.* Pay for Performance has been based on Denver's unprecedented collaboration between the Board of Education and the Association. These parties have used their dual commitment to student achievement as a basis for sponsoring and regularly improving a high-risk venture. Their collaboration has proven instrumental to making mid-course corrections that have consistently strengthened Pay for Performance. In so doing, they have demonstrated a different way of conducting business on behalf of students. This collaboration has been substantive and effective. It should be extended to other parts of district educational operations, regardless of the outcome of the Association and Board votes on a new compensation plan.
- *Continue to place problems on center stage.* A central factor contributing to the accomplishments of the pilot has been the ability to place critical issues that affect the district on center stage. Urban school districts are characteristically reluctant to make their most serious internal problems highly visible. Yet doing so has been a major strength of Pay for Performance. Operating in a climate protected by external supporters and internal reformers, the pilot provided a functional vehicle for multiple problems to be identified, discussed and then acted upon. The district will benefit by continuing and extending this function.
- *Create a Principals Leadership and Achievement Institute.* High quality principals are key to shaping, guiding and supporting school improvements. Under Pay for Performance, the district's new educational initiatives and No Child Left Behind, their responsibilities are expanding and their decisions are becoming more critical to the success of students and publicly visible. All principals need to understand deeply how learning occurs and how it is nourished, measured and supported. They need ongoing, sustained opportunities to identify salient site issues, analyze trends in student achievement data, reflect on emerging issues, develop their skill in observing classrooms and providing support to teachers, and build the knowledge to work effectively with diverse students and families. These functions are at the core of thoughtful and anticipatory school leadership. Building these capacities will complement the current district plans to prepare principals to carry out targeted educational initiatives, a Principals Leadership and Achievement Institute will provide the vehicle needed for strengthening these abilities in Denver's principals.
- *Prepare for the post-pilot and post-vote transition.* The pilot benefited greatly from having a special internal implementation team with the commitment and sense of urgency that is essential to create change. As the learnings and practices from the pilot are implemented district-wide, it is now essential to institutionalize these qualities. The supports for the new compensation plan and expanding educational initiatives need to be embraced by and channeled through the formal district structures. This is a critical step for the district even with the Design Team on board for the next phase of implementation. Change agents can function effectively within large bureaucracies; however, they are not bureaucrats. As the transition from the "old Denver Public Schools" to the "new Denver Public Schools" continues, the district needs to ensure that the departments and units of the system are functioning with increased capacity, accountability and urgency on behalf of the district's educational initiatives.

IMPACT

At school and central levels, the role of leadership in a school district is to look ahead, anticipate the needs of students, and create new approaches and

solutions to existing and emerging educational problems. It takes courage, integrity, and personal accountability to teach all students in a district as diverse as Denver. The pilot has revealed outstanding leaders; it has also revealed gaps in leadership knowledge and skill and vacuums in leadership. Due to the accomplishments of the pilot sponsors, the new administration, the Design Team and the Joint Task Force, Denver now has a pivotal opportunity for leaders to expand a new way of conducting business.

C. Summary

The Board of Education and the Denver Classroom Teachers Association have moved the Denver Public Schools to the forefront of educational reform in the United States. Moreover, the parties have committed to studying the pilot and regularly making results available to local and national audiences. Rather than introduce a piecemeal reform, they have sponsored and supported Pay for Performance as it has moved the entire organization to make improvements which help students to learn and teachers to be more effective in the classroom.

The pilot has demonstrated that a focus on student achievement and a teacher's contribution to such achievement can have a far reaching institutional effect—*if the initiative also addresses the district factors that shape the schools.* In so doing, Pay for Performance has further shown that the issue of organizational alignment cuts to the very essence of how—and to what extent—a school district is functioning in support of student learning. The challenge ahead for the district is to build on the pilot's foundation when implementing next iterations of Pay for Performance, undertaking the district's other educational initiatives, and meeting the requirements of No Child Left Behind. Pay for Performance has been a catalyst for change.

CHAPTER X

National Implications

A. Introduction

Linking what teachers earn to what students learn can be a major lever in support of fundamental systemwide change in school districts. Pay for performance—when well implemented—has the salutary effect of forcing a *district* to operate in a much more effective and efficient fashion in support of student learning. That is to say, changes in district practices that are necessary to advance pay for performance also directly support quality teaching and enhanced learning.

Based on this premise, the following discussion has four purposes. First, it lays out the core considerations for districts when undertaking pay for performance initiatives. Second, it identifies the types of assistance districts characteristically require to redefine traditional roles, practices, and policies. Third, the chapter examines learnings for private foundations that have emerged from the pilot. Lastly, it presents philanthropic strategies for extending the potential and reach of pay for performance.

B. Core Considerations

Support and accountability are the twin pillars of sustainable reform in school districts. Embracing either one, to the exclusion of the other, is essentially selecting one form of myopia over another. The potential power of pay for performance is in focusing on both support and accountability. It therefore can be integral to critical reforms in public education. To be successful, though, districts need to learn from the failed attempts of the past and to overcome the skepticism and barriers related to tying individual teacher performance legitimately to student achievement. For districts preparing to head in the direction of pay for performance, the following considerations can be the keys to success.

Process

Pay for performance functions best when it reinforces a district's core goals; it is not a freestanding program or an adoptable model. Accordingly, the basic elements needed to undergird any customized, systemic reform have to be considered when launching a major district initiative which links student achievement to compensation. These elements include:

- *Collaboration.* Providing substantive opportunities for teachers and principals—not only their leaders—to shape, steer and influence the initiative refines process and strengthens the outcome. Collaboration must be present from the start of the effort through all phases of the design, development and implementation. Simply put, pay for performance imposed by a board or district leadership erodes the potential to develop real accountability.
- *Trust.* A high level of trust is required for any effort that seeks to link student achievement, adult performance and evaluation, and compensation. Participants need to be convinced that the initiative is intended to be supportive of teachers, rather than punitive. Therefore, on the front end, the initiative needs to build trust among diverse constituencies. This includes the trust between the schools and the district, between principals and teachers, and between and among teachers.
- *Communication.* Major initiatives are frequently derailed by gaps in information and communication. Indeed, in the field of public education, the forces of misinformation are often greater than the forces of accurate information. In an era when accountability often takes the form of a hammer on perceived underperformers, it is essential to craft, carry out, regularly review and strengthen a communications strategy.
- *Phases.* Pay for performance is a marked departure in culture and practice for school districts. During implementation, it will stretch the support capacities of a district. It should be introduced in phases so that the district will have the opportunity to make mid-course corrections and improvements as necessary. Otherwise, the distinction between the intent of the initiative and how it is being implemented will get lost. When this occurs, participants will blame and subsequently oppose the initiative; this is a recurring pattern over many years in American education.
- *Relation to Collective Bargaining.* Teacher unions are taking leadership in the performance pay arena and their commitment to the design and implementation of an initiative is essential. However, during the developmental phases, the initiative should be discrete and separate from the negotiation process. When implemented thoughtfully, pay for performance focuses on core conditions affecting teaching and student achievement. Collective bargaining focuses on working conditions. If the two dovetail too quickly the confidence in pay for performance will be undermined.

Purpose

From the outset, it is essential to be clear on the purpose of the initiative; this significantly affects the results. With many performance-based initiatives, multiple purposes often compete, pulling the initiative in different directions. For example, the goals of building a new compensation plan or changing professional development may sometimes work against the goal of improving achievement. What then occurs is that the focus on improving student achievement becomes blurred or merged with other purposes, leading to confusion on the part of teachers and administrators and competition among district priorities. There needs to be real clarity on what will be rewarded and why.

Link of Student Achievement to Compensation

When the primary purpose is to improve student achievement, the initiative becomes easier to understand, implement, support and evaluate. As just one example, if the purpose is *increasing* student achievement, the clear tie to the delivery of instruction and to motivating students becomes vitally important. Then, the need to provide sustained support to classroom teachers becomes paramount. When this is not the primary purpose

of the initiative, the entire emphasis on student learning can become muddied or lost. Then, pay for performance often deteriorates into a failed effort to create incentives for teachers. Student achievement needs to be both the driver and end result; this cannot be overemphasized.

Data and Assessment Capacity

Pay for performance puts new demands on teachers. For example, it demands that they pay attention to the starting places of each of their students in various subjects—that they study the data and understand each student's status—and that they build lessons and interventions based on this knowledge. The specific identification of each student's status at the beginning and end of the school year, and over multiple years, is required for the purpose of measuring the results. An understanding of student academic progress is required for the teacher to develop appropriate lessons.

Such requirements of teachers, in turn, place demands on the district that may be surprisingly difficult to meet. If teachers are to work with data, for example, they must have that data available to them at the beginning of school in a form that is timely, usable and complete. In most districts, this has not been a requirement in the past. Even in the current era of No Child Left Behind, most districts are initially unable to meet this demand. As this problem is addressed, it helps advance pay for performance, while also helping all schools and all teachers. This kind of data capacity in support of pay for performance is critical for its success.

Assessment is necessarily at the core of any pay for performance initiative, as it is for much of the school improvement and accountability efforts being attempted across the nation. Indeed, the requirements of an assessment system under pay for performance are essentially the same as for implementing No Child Left Behind or any effort which seeks to link student and teacher performance.

The potential of performance-based initiatives can be undercut if assessment-related issues are overlooked. Too frequently, the purposes of the assessments are unclear, assessment results are inaccurate, or the interpretation applied to test results exceeds what may legitimately be inferred from those results. These problems are serious enough when assessments carry high stakes for students. When you add teacher compensation to the stakes, the need for reliable assessments—fairly constructed and accurately interpreted—becomes critical.

Several key considerations regarding assessment are indicated below. This listing is not intended to be all-inclusive. Rather, it highlights pivotal challenges which can be addressed and which should not be allowed to serve as barriers to undertaking pay for performance.

1. *Student Growth.* In a pay for performance system, a district must base its evaluations of teacher performance, in part, on student growth. Therefore, its assessments must measure individual student growth. Many state assessments speak of growth, but are used to compare one group of students—one class of 4th or 7th graders—with the previous year's class. While this kind of assessment often provides valuable information, it compares different groups of students and does not reflect the growth of individual students. To the extent that one group of 25-30 students differs from another, which can be considerable, these groups cannot fairly be compared to each other, and do not describe the success of a teacher with his or her class of students.
2. *Baseline Data.* Measuring student growth assumes a starting point and an ending point. A student's reading level at the end of fourth grade may be an absolute, but without knowing that student's prior reading level it is not possible to infer from a single score how much the student's reading has improved or what has been the contribution of that student's teacher. For this reason, there must be baseline data for each student, as well as for any broader comparisons that are undertaken.
3. *Link to Curriculum and Instruction.* If an assessment does not measure what was taught, it cannot be said to reflect a teacher's contribution to what was learned. Thus, assessments that may be generally useful in gauging student knowledge may not be useful measures of teacher effectiveness. Similarly, if the teacher does

not teach to the curriculum, even an assessment aligned to the curriculum does not measure teacher effectiveness. The latter is an issue that would have to be addressed administratively.

Validity

There are three kinds of validity that pay for performance—and any program measurement—must address. First, there is statistical or scientific validity. Whatever measures are reported or actions taken should be the result of assessments that are measured using statistically valid methods. While this point seems obvious, many states, districts and even test companies fail the test of statistical validity in the inferences they draw from their tests. Statistical validity is difficult to achieve at the classroom level, since the numbers of students are small and the possibility that a variation may be attributable to chance or aberrant scores is correspondingly great. There are various approaches to addressing this problem, including using multiple measures of achievement and/or multiple years of a teacher's results. While these methods add some complexities to the process, they can be used to increase the statistical validity of an assessment, making it both fairer and more useful.

Second, there is educational validity. It is possible for statistical results to support practices that are not educationally valid, at least in the short term. It is also quite possible for educationally sound practices to be difficult to measure or prove statistically. Any initiative put into place must also satisfy what is known about how students learn: it must have educational validity.

Third, there is political validity. This becomes extremely important if comparing scores on standardized tests is one of the methods being used to gauge teacher success. Even where results are significant statistically, they may not be perceived as legitimate. If teachers perceive that measures being used to partially determine their compensation levels are not legitimate, no amount of statistical validation will be of value. Therefore, political validity—the *perception* that the system is fair—is critically important at every step of the development process.

Organizational Alignment

If the purpose of the initiative is increasing student achievement, the organization must align itself around that goal in a much clearer and more effective manner than is often the case in school districts.

A pay for performance system demands that a district's standards, curriculum content, instructional delivery, professional development, data capacity, assessment, supervisory and human resources be aligned. This is frequently not the case. Numerous failed reforms nationally have been based on the notion that single components of a district's practices can be altered in ways that will change the rest of the system. However, the issue of alignment reaches far beyond organizational structures or the currently popular intervention of the day. It cuts to the very essence of how—and to what extent—the school district is functioning systematically in support of student learning. Addressing the issues of organizational alignment is pivotal to the initiative's prospects for success.

Professional Development

Professional development is a critical component of successful change. In a pay for performance plan, it is also critical to the success of the plan itself. The expectations of pay for performance include that teachers and principals obtain student achievement data, analyze the results, and tailor instruction both to the curriculum provided and the students' abilities and needs. This requires that the teachers and principals have the appropriate data available, and that they are able to understand and interpret the data accurately, identify student needs, set appropriate learning objectives, and structure lessons accordingly. In our experience, however, even excellent teachers may not have all of these skills, particularly those relating to data. To fairly gauge a teacher's instructional ability, therefore, professional development is required.

Professional development may also be required in standards-based instruction, differentiated instruction, or other related skills. Providing professional development in each of these areas enhances and reinforces the fairness, effectiveness and accuracy of the pay for performance initiative.

It also increases the likelihood of increasing student learning, by addressing critical gaps in the instructional process.

Costs

The range of costs connected to implementing pay for performance initiatives and making systemic changes take two forms. First, there are costs that are financial in nature. These result from new fiscal outlays such as salaries, equipment and additional staffing. School boards, unions and superintendents are highly familiar with the financial costs of change. Second, there are costs related to changing practices. These are non-financial in nature and frequently underestimated. They include the institutional costs of reordering district priorities, functioning with higher levels of inter-departmental coordination, operating with a greater sense of urgency and reallocating existing funds. With an organizational priority 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.

C. Services and Assistance

An extensive range of capacities is required to implement pay for performance effectively. The problem, though, is that most districts lack this breadth of capacity. As a result, there is a repeated national pattern of district misfires as they launch new initiatives. This is not a function of poor intent. Rather, it is because districts need assistance to develop new capacities as they plan and implement major initiatives. Absent such support, the already serious challenge of implementing pay for performance is exacerbated.

Districts require assistance in the multiple phases of conceptualizing, developing, implementing and evaluating a pay for performance initiative. The following highlights several of the substantive areas in which responsive technical assistance can markedly increase district capacity. This listing is representative, rather than all-inclusive.

Identify initial levels of readiness and capacity

An essential, often overlooked step in preparing for pay for performance is to conduct a district assessment. This includes building the base needed to:

- Identify key participants for the buy-in, design, implementation and policy approval phases.
- Assess the district's current level of readiness to pursue a pathway of fundamental reform.
- Determine the district's current performance and capacity in the areas needed to support and implement pay for performance.
- Assess the current process for teacher evaluation.
- Determine the district's ability to link student and teacher data.
- Identify the key constraints—legal, cultural, district rules and policies, existing contracts—that may affect the prospects for pay for performance.

Customize the design and implementation strategy for pay for performance

Pay for performance needs to be approached systematically. However, many districts have gone down this path with a series of tactics, but in the absence of a strategy. As a result, they have lacked the ability to be anticipatory and to overcome obstacles that emerge during implementation. Assistance is needed to:

- Develop a district-appropriate definition of pay for performance.
- Align this initiative with district goals.
- Determine the structures and participation necessary to design and implement the initiative.
- Introduce and support pay for performance as a vehicle for promoting and supporting improved student achievement and quality teaching.
- Define project plans and phases, targets, resource requirements and timeframes.
- Establish project management goals, benchmarks and reports.

- Define and communicate the accountability mechanisms.
- Secure requisite internal and external resources.

Build the base of institutional, constituent and community support

Pay for performance requires a broad base of support, both within and from outside the district. Internally, it can only succeed with significant buy-in from teachers and principals. It also requires commitment from the central administration. Moreover, it must be one of the highest priorities of the superintendent, the school board and the teachers union. Externally, community and corporate support are necessary, both to help overcome entrenchments within the district and to provide long-term financial support. Eventually, it must be approved by teachers throughout the district and by the school board.

Building this base of support requires the ability to conduct consistent, sophisticated communications and community organizing. This, in turn, necessitates assistance that develops the capacities to:

- Prepare and implement a coordinated communications strategy.
- Build a district- and community-wide understanding of pay for performance.
- Provide outreach to external grassroots and institutional constituencies, and the media.
- Train constituent groups (board members, teachers, site administrators, union officials, central administrators, parents, community members and other external parties) in understanding the design and implementation phases of the initiative.
- Provide avenues for ongoing constituent input, participation and response.
- Demonstrate the improvements in learning, teaching and organizational support resulting from the initiative.
- Respond rapidly to clarify any major areas of confusion regarding the initiative.

Strengthen district data capacity

A critical challenge when implementing a compensation system based on student achievement is to determine the extent of learning and progress district-wide, school by school, classroom by classroom, and student by student. The district needs to know which students are succeeding, which students are not succeeding, and why. This knowledge is essential for realigning district resources based on the needs of children at each individual school site and for establishing expectations to which everyone in the district will be held accountable.

Assistance is needed to train key staff in the development of a comprehensive district accountability system. This specifically focuses on building the capacities to:

- Identify actual and relative school performance.
- Disaggregate district performance indicators by various student-related subgroups such as socioeconomic status, race, ethnicity, mobility, etc.
- Identify the student subgroups that have the greatest needs and represent the greatest opportunity for improvement—e.g., those whose performance is substantially below that of the best performing subgroup.
- Calculate the performance of the various groups at select schools through a process that can then be applied to all schools.
- Analyze similarities in results among high-performing schools and the differences in results between the high and low performing schools.
- Disaggregate data by grade and classroom to provide comprehensive, multi-year individual student growth data to teachers.
- Focus on longitudinal analysis based on individual student growth.
- Involve principals, teachers and parents in developing the data presentation formats.

Design the compensation plan

This focuses on all aspects of developing, field-testing, finalizing and engendering support for a

new district direction for compensation. It includes building the capacities to:

- Evaluate the current system of salaries and benefits.
- Examine different kinds of compensation systems within both the corporate and public educational sectors.
- Differentiate between the myths and realities of such systems.
- Evaluate the impact of different approaches to teacher compensation on student achievement.
- Identify relevant, effective practices.
- Craft a customized, pay for performance component in the compensation system.
- Assess the plan's financial implications.
- Determine the vehicles for making the transition from the existing, negotiated salary schedule to the new plan.
- Build appropriate expectations within the district.
- Determine the strategy for field-testing and improving the new design.
- Ensure that the implementation of the plan can withstand leadership changes.

Build leadership and organizational alignment

An expanded base of leadership is needed to develop, implement and make mid-course corrections to the pay for performance initiative. Leadership is particularly needed—at school, district and policy levels—to ensure that the school district is aligned in support of pay for performance.

This necessitates building the capacities to:

- Analyze and strengthen the alignment between school and classroom goals, curriculum content, the planning and delivery of instruction, assessments, professional development and compensation.
- Integrate and upgrade the student information and human resources data systems.
- Identify and address gaps in the existing supervisory and support structures.
- Incorporate student achievement trend analysis into board program and policy decision making.
- Link the emerging pay for performance initiative with the requirements of the No Child Left Behind Act.
- Create new expectations for performance throughout the district.
- Redefine roles and clarify changes in responsibility.
- Evaluate leaders using measures which include student achievement.
- Train principals and central administrators in the development and interpretation of individual and school level student achievement data focusing on the growth of individual students and school trends.
- Integrate and coordinate the systems for teacher evaluation, support and recognition with student achievement being both the driver and end result.

Establish a comprehensive professional development strategy

The precursor to such a strategy is to conduct a rigorous professional development audit. This will provide a detailed analysis necessary for reallocating and deploying existing resources. The components of the audit should include: (1) defining the initiating events for the professional development, (2) detailing training offerings, (3) assessing the content of the training, (4) assessing the frequency of the training, (5) identifying the recipients of the training, by session, (6) identifying the subsidy source, (7) identifying the subsidy amounts, (8) assessing the providers and the number of staff involved, (9) examining the providers in terms of their placement within the organization, (10) reviewing quality indicators and determinations, (11) evaluating the level of mastery demonstrated by those trained, and (12) determining the overall impact of the professional development provided.

This information then becomes the basis for preparing the district's professional development strategy and aligning it with instructional goals. Using this approach, the district is better able to provide school sites with professional development based on actual student achievement results and the differentiated needs of the school sites—a sound educational practice. Moreover, particularly during a period of fiscal austerity, it enables a district to better target resources to improve student achievement.

D. Foundations

When seeking to invest in public schools, foundations have often followed the pattern of making incentive grants available for purposes of planning and implementation. The applicants characteristically must follow foundation-defined templates and priorities. Initially, a modest number of schools or districts secures funds for a time-limited planning period. Subsequently, they apply for larger grants; the foundation then selects a smaller number of these schools or districts for a multi-year implementation period.

As a recurring approach to educational philanthropy, particularly at the national level in recent decades, it characteristically produces a dynamic that generally falls short of the intended results. It encourages short-term responses to a grant incentive, rather than fundamental change. When the extra funds and special dispensations are no longer available, the system returns to earlier patterns of practice and there are few resultant learnings. Various philanthropic requirements such as requiring matching funds or embedding new concepts in contracts have not changed this basic outcome. Simply put, the way the system thinks and behaves does not change.

A new form of philanthropy more wisely follows a different path. It is sensitive to the need to not impose foundation priorities from above. Yet it is also careful because pursuing a primarily bottom-up approach carries with it several cautionary red flags and distinct gaps, including the lack of district commitment to the initiative. Supporting new ventures and advocating for fundamental change requires a creative, concurrent top-down, bottom-up philanthropic strategy.

Venture Capital

Achieving different results requires new ideas whose implementation can be tested and critically examined. This, in turn, can only be achieved with a different approach to philanthropy. With philanthropic leadership, a significant change was advanced in Denver. Rather than invest in the model program currently in vogue or a foundation-created construct, a blend of local and national funders invested in a far-reaching district and union experiment with a concept, pay for performance.

This is one of the rare examples of foundations applying the approach of venture capital to public educational giving. The foundations took significant risks in supporting the field testing and study of an unproven venture linking teacher compensation, in part, to student achievement. While the potential for district learning and change was great, so too was the potential for public embarrassment for the foundations. Embarking upon this direction required leadership—a critical element for achieving philanthropic impact

Throughout the pilot, the foundations sustained their giving, while concurrently broadening their own knowledge as well as those of the district. In describing the philanthropic community and the district, one foundation executive noted, "I don't think any of us, including the administration, really correctly estimated the size or amount of work entailed in this project." Another added, "We all have learned that this subject is a lot more complicated than we first thought. The system has also learned a lot from this effort."

The definition of success also expanded as the pilot achieved greater reach into the system and encountered barriers to progress. An executive director of a foundation commented, "Success is not just a blanket commitment to Pay for Performance. Success is measured by how much learning occurs and whether the learning is used to make positive changes."

Results: Research as Driver

The venture capital support was triple-tiered, supporting direct services (e.g., communications), a range of technical assistance, and the

research study. The latter was particularly stressed due to the philanthropic emphasis on having district actions be based on learnings. A leader of a local foundation states, "Regardless whether [a new compensation] plan passes or not, Pay for Performance will have an impact. The vote will not be the final word.... [the foundation] gave the money so that district leadership could learn. They need to get beyond the usual inertia."

Just as pay for performance focuses on results, the foundations also stressed results. When research produced findings and recommendations, the foundations wanted to see follow-up action. This, in turn, helped internal district reformers to introduce changes. Another foundation leader comments, "In this field, no distinction is made between an educational concept and the execution of the concept.... What to do in education is up for grabs. The impact of this project is important and designates a new time and age.... The challenging of district personnel and the system as a whole is reshaping the district."

It requires a delicate balance to push grantees and their beneficiaries for results, yet avoid functioning as de facto operating foundations. Particularly in the area of national school reform, this quandary has been problematic for many foundations. Consequently, they have created accountability and partnership mechanisms which grantees are characteristically required to use. Beyond their varying levels of effectiveness, these have often been received as imposed mechanisms. Differing from this approach, the Pay for Performance supporters urged parties to collaborate, joined in the collaboration when asked and provided additional support when gaps or deficiencies in the pilot were revealed. However, the mechanisms for leadership and accountability were neither created nor imposed by the foundations.

Establishing the Context

Districts exist in a larger context—equally as political as educational—which needs to be influenced and shaped to create an environment that is more open and conducive to a change as significant and far-reaching as pay for performance. Indeed, in our meetings with policy makers, districts, funders, commentators and the media, it is notable how opinionated most parties are on the topic of pay for performance—regardless of their base of factual information. It is reminiscent of the old axiom that it is incredible what conclusions you can draw if you do not bother to let the facts get in the way. This has serious implications for experiments with pay for performance.

Foundations need to take an expansive approach consistent with their core belief in the fundamental importance of improving public education. The need is to create a local and/or national context in which trailblazing districts and unions can explore and experiment with needed avenues for major change. This will build on a foundation's rightful goal of supporting local initiatives (without imposing its priorities on the field), while shaping the context for Pay for Performance or comparable initiatives in ways that will enhance the prospects for significant—and critically needed—success.

Recognizing there is controversy surrounding and opposition to these issues, foundations can use their pivotal philanthropic role to help re-shape the context, climate and discussion of performance-based improvement strategies. By creating safe havens for disagreeing or conflicting parties, highlighting the need for actions in response to research findings, and being committed for the long haul, foundations can greatly expand the reach of their philanthropy and the impact of systemic initiatives such as Pay for Performance.

Differing dramatically from traditional educational philanthropy, this approach is rooted in venture capital. It focuses on fundamentally changing conditions and enabling school districts to move in new directions. This directly supports the goal of ensuring that students—and those who contribute to their achievement—are the direct beneficiaries of the improvement efforts.

E. Summary

There are a range of the factors that must be considered when developing and implementing pay for performance initiatives. Such efforts can result in a new approach to rewarding teachers, whether that consists of small bonuses or a large restructuring of the compensation system. Most significantly though, these initiatives can be a catalyst for aligning district resources, actions and expectations in

support of the overall goal of increasing student achievement and supporting teachers. In this way, pay for performance can provide a basis for improving the entire school system by tying district activities to core classroom needs. When the school *system* is functioning in a much more organized and effective manner in support of better teaching and enhanced learning, pay for performance can become a vehicle for increasing student achievement—*the* bottom line for school reform.

Endnotes

Chapter I

1 Kohn, A., The Folly of Merit Pay, *Education Week* (September 17, 2003).

2 "Rewarding Teacher Quality," available at http://www.nga.org/incentivepay/ [2002, November].

3 From an internal project planning document of the Design Team, December 3, 1999.

4 *Pathway to Results* can be found on the website of the Community Training and Assistance Center at www.ctacusa.com.

Chapter II

1 Sanders, W. J. and Horn, S. P. (1998). Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research. *Journal of Personnel Evaluation in Education,* 12(3), 247-256.

2 Haycock, K. (1998). Good Teaching Matters . . . A Lot. *Thinking K-16.* 3(2), 3-14.

3 Wilms, W. and Chapleau, R. (1999, November 3). The Illusion of Paying Teachers for Student Performance—Some Lessons from History. *Education Week.* 19(10) 48, 34.

4 Moore Johnson, S. (1986). Incentives for Teachers: What Motivates, What Matters. *Educational Administration Quarterly,* 22(3), 54-79.

5 Odden, A. (2000, January). New and Better Forms of Teacher Composition are Possible, *Phi Delta Kappan,* 361-366.

6 Archer, J. (2001, February 7). Business Seeks Teacher Renaissance, *Education Week,* 34, 48.

7 Archer, J. (2000, June 21). NEA Poised to Debate Pay for Performance. *Education Week,* 5.

8 Merriam, S. B. (1988). *Case Study Research in Education: A Qualitative Approach.* San Francisco: Jossey-Bass.

Bogdan, R. C. and Knopp Biklen, S. (1982). *Qualitative Research for Education: An Introduction to Theory and Methods.* Boston: Allyn and Bacon.

Yin, R. K. (1984). *Case Study Research: Design and Methods.* Beverly Hills: Sage.

Chapter III

1 Previously cited Agreement, Appendix E, Currently Appendix 01-01(1), III-A-3.

2 This is a math instrument developed by the district for use in the Title I program, which was later renamed Grade Level Math and is used in the district for Pay for Performance.

3 Denver Public Schools (2000). *District Assessment Guide: Elementary and Secondary.* Denver, CO.

Chapter IV

1 See links for teacher lesson planning at the following: http://www.askeric.org/Virtual/lessons/

2 Mager, R. (1962). *Preparing Instructional Objectives.* Palo Alto, CA: Fearon.

3 Fraser, B. J. et al. (1987). Synthesis of Educational Productivity Research, *Journal of Educational Research,* 11, 2, 145-252.

4 Wolfe, P. (Nov 1998). Revising Effective Teaching, *Educational Leadership,* V56, N3, 61-64.

5 Walberg, H. J. (1999). Productive Teaching, in Waxman, H.C. and Walberg, H.J. (Eds.), *New Directions for Teaching Practice and Research,* Berkeley, CA: McCutchen, 75-104.

Chapter V

1 Darling-Hammond, L. (2000, January). Teacher Quality and Student Achievement: A Review of State Policy Evidence, *Education Policy Analysis Archives* (8:1), pp. 3-6 at http://epaa.asu.edu/

2 Hinde, E. (2003). Reflections on Reform: A Former Teacher Looks at School Change and the Factors that Shape It, *Teachers College Record* at http://www.tcrecord.org/ ID Number 1183, published 8/3/2003.

Chapter VIII

1 Bryk, A. S. and Schneider, B. L. (2002). *Trust in Schools: A Core Resource for Improvement*. Russell Sage Foundation.

2 Southern Regional Education Board (SREB) (2002). *Quality Teachers: Can Incentive Policies Make a Difference?* 13.

3 Henderson-Montero, D., Julian, M. W., and Yen, W. M. (2003, Summer). Multiple Measures: Alternative Design and Analysis Models. *Educational Measurement: Issues and Practice*, 22, 7-12.

Notes

Appendix

Figure A-1 through A-12, A-43 through A-46 The second and third models adjust for the following school factors: principal years at the school, percent of students with disabilities, percent of students who are English language learners, percent of students receiving free or reduced price lunch, percent of teachers not fully licensed, and total school enrollment. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-13 The second and third models adjust for the following school factors: percent of students with disabilities, percent of students who are English language learners, percent of students receiving free or reduced price lunch and total school enrollment. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-14 The second and third models adjust for the following school factors: principal years at the school and percent of students with disabilities. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-15 The second and third models adjust for the following school factors: percent of students with disabilities and percent of students receiving free or reduced price lunch. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-16 through A-18 The second and third models adjust for the following school factors: principal years at the school, percent of students with disabilities, percent of students who are English language learners, percent of students receiving free or reduced price lunch and total school enrollment. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-20 The second and third models adjust for the following school factors: percent of students with disabilities, percent of students receiving free or reduced price lunch and total enrollment. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-21 The second and third models adjust for the following school factors: principal years at the school, percent of students who are English language learners, percent of students receiving free or reduced price lunch, percent of teachers not fully licensed, and total school enrollment. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-22 The second and third models adjust for the following school factors: percent of students with disabilities, percent of students who are English language learners, percent of students receiving free or reduced price lunch, and total school enrollment. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-23 through A-24 The second and third models adjust for the following school factors: percent of students with disabilities and percent of students who are English language learners. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-25 The second and third models adjust for the following school factors: percent of students receiving free or reduced price lunch and percent of teachers not fully licensed. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-26 through A-27, A-29 The second and third models adjust for the following school factors: principal years at the school and percent of students with disabilities. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-28, A-30, A-31 The second and third models adjust for the following school factor: percent of students receiving free or reduced price lunch. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-32 through A-42, A-55 through A-64 The second model adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

Figure A-49 through A-54 The second and third models adjust for the following school factor: percent of students with disabilities. The third model also adjusts for the following student factors: low SES, any disability, retained a grade, English proficiency, ethnicity, and gender.

FIG. A-1

PFP Effect—Elementary Schools—*ITBS* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	43.5	0.0001	42.3	0.0001	56.1	0.0001
Pilot	0.4	0.7894	2.1	0.0156	2.5	0.0024
Control	0.0		0.0		0.0	
Time	-0.3	0.0004	-0.1	0.2870	0.1	0.0847
Time x Pilot	0.2	0.1374	-0.1	0.5716	-0.3	0.0318
Time x Control	0.0		0.0		0.0	
Number of Observations	49592		49592		49592	

FIG. A-2

PFP Effect—Elementary Schools—*ITBS* Language NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	43.7	0.0001	43.5	0.0001	56.5	0.0001
Pilot	-2.3	0.1027	-0.7	0.4885	-0.6	0.5398
Control	0.0		0.0		0.0	
Time	-0.2	0.0205	-0.1	0.1262	-0.03	0.6427
Time x Pilot	-0.1	0.5093	-0.1	0.6769	-0.1	0.6696
Time x Control	0.0		0.0		0.0	
Number of Observations	44486		44486		44486	

FIG. A-3

PFP Effect—Elementary Schools—*ITBS* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	43.6	0.0001	42.7	0.0001	52.6	0.0001
Pilot	2.3	0.1088	3.3	0.0003	3.7	0.0001
Control	0.0		0.0		0.0	
Time	-0.6	0.0001	-0.4	0.0001	-0.3	0.0001
Time x Pilot	-0.1	0.5583	-0.2	0.2325	-0.4	0.0128
Time x Control	0.0		0.0		0.0	
Number of Observations	47164		47164		47164	

FIG. A-4

PFP Effect—Elementary Schools—*CSAP* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	51.6	0.0001	51.0	0.0001	65.3	0.0001
Pilot	2.9	0.0645	2.1	0.0289	2.6	0.0018
Control	0.0		0.0		0.0	
Time	-0.2	0.0735	-0.01	0.8988	0.2	0.0516
Time x Pilot	-0.3	0.0694	-0.5	0.0072	-0.7	0.0001
Time x Control	0.0		0.0		0.0	
Number of Observations	36398		36398		36398	

FIG. A-5

PFP Effect—Elementary Schools—*CSAP* Writing NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	49.5	0.0001	49.4	0.0001	63.0	0.0001
Pilot	6.1	0.0010	3.2	0.0018	3.7	0.0001
Control	0.0		0.0		0.0	
Time	0.1	0.2053	0.3	0.0041	0.5	0.0001
Time x Pilot	-0.5	0.0220	-0.7	0.0025	-0.8	0.0001
Time x Control	0.0		0.0		0.0	
Number of Observations	24463		24463		24463	

FIG. A-6

PFP Effect—Elementary Schools—*CSAP* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	51.0	0.0001	49.4	0.0001	61.2	0.0001
Pilot	5.5	0.0253	5.2	0.0002	6.2	0.0001
Control	0.0		0.0		0.0	
Time	0.002	0.9863	0.2	0.2730	0.2	0.1618
Time x Pilot	-1.3	0.0001	-1.3	0.0001	-1.5	0.0001
Time x Control	0.0		0.0		0.0	
Number of Observations	11154		11154		11154	

FIG. A-7

PFP Effect—Middle Schools—*ITBS* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	43.1	0.0001	42.6	0.0001	55.6	0.0001
Pilot	-10.3	0.1445	-2.4	0.3592	-2.9	0.3047
Control	0.0		0.0		0.0	
Time	-1.1	0.0001	-1.3	0.0001	-0.4	0.0001
Time x Pilot	0.8	0.0029	0.9	0.0024	1.1	0.0001
Time x Control	0.0		0.0		0.0	
Number of Observations	43375		43375		43371	

FIG. A-8

PFP Effect—Middle Schools—*ITBS* Language NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	46.4	0.0001	46.1	0.0001	57.7	0.0001
Pilot	-6.7	0.2628	-3.3	0.3148	-2.8	0.3326
Control	0.0		0.0		0.0	
Time	-1.0	0.0001	-1.1	0.0001	-0.4	0.0001
Time x Pilot	-0.3	0.3006	-0.4	0.1394	-0.3	0.1982
Time x Control	0.0		0.0		0.0	
Number of Observations	41493		41493		41490	

FIG. A-9

PFP Effect—Middle Schools—*ITBS* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	42.1	0.0001	41.4	0.0001	51.4	0.0001
Pilot	-7.2	0.2528	-0.9	0.6854	-1.5	0.5598
Control	0.0		0.0		0.0	
Time	-0.8	0.0001	-1.0	0.0001	-0.2	0.0238
Time x Pilot	0.3	0.2103	0.4	0.2163	0.3	0.1828
Time x Control	0.0		0.0		0.0	
Number of Observations	41815		41815		41812	

FIG.A-10

PFP Effect—Middle Schools—*CSAP* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	49.5	0.0001	49.0	0.0001	63.5	0.0001
Pilot	-8.9	0.2199	0.5	0.8232	0.9	0.6991
Control	0.0		0.0		0.0	
Time	0.03	0.7794	-0.3	0.0149	0.5	0.0001
Time x Pilot	0.3	0.3765	0.1	0.8407	-0.1	0.8075
Time x Control	0.0		0.0		0.0	
Number of Observations	38363		38363		38359	

FIG. A-11

PFP Effect—Middle Schools—*CSAP* Writing NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	49.7	0.0001	49.5	0.0001	64.2	0.0001
Pilot	-9.0	0.2093	-0.9	0.6476	-1.1	0.6486
Control	0.0		0.0		0.0	
Time	0.003	0.9800	-0.3	0.0080	0.6	0.0001
Time x Pilot	0.3	0.3374	0.5	0.1843	0.7	0.0455
Time x Control	0.0		0.0		0.0	
Number of Observations	30201		30201		30197	

FIG. A-12

PFP Effect—Middle Schools—*CSAP* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	50.0	0.0001	49.6	0.0001	59.9	0.0001
Pilot	-8.9	0.1887	-0.9	0.7356	-1.3	0.5782
Control	0.0		0.0		0.0	
Time	-0.3	0.0187	-0.6	0.0001	0.6	0.0001
Time x Pilot	1.9	0.0001	1.9	0.0001	1.6	0.0001
Time x Control	0.0		0.0		0.0	
Number of Observations	30279		30279		30275	

FIG. A-13

PFP Effect—High Schools—*ITBS* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	45.2	0.0001	44.6	0.0001	55.5	0.0001
Pilot : Manual	-13.0	0.0601	2.2	0.6954	-4.3	0.4486
Pilot : Jefferson	10.7	0.3246	-3.1	0.6434	-2.4	0.7261
Control	0.0		0.0		0.0	
Time	0.9	0.0001	1.4	0.0001	2.0	0.0001
Time x Manual	-0.2	0.8712	-0.01	0.9941	1.9	0.0912
Time x Jefferson	-2.0	0.0023	-2.1	0.0025	-1.6	0.0113
Time x Control	0.0		0.0		0.0	
Number of Observations	20000		20000		19995	

FIG. A-14

PFP Effect—High Schools—*ITBS* Language NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	35.7	0.0001	34.1	0.0001	43.7	0.0001
Pilot : Manual	-13.9	0.0052	-18.8	0.0022	-28.0	0.0026
Pilot : Jefferson	-9.7	0.1626	-13.3	0.0927	-17.5	0.1672
Control	0.0		0.0		0.0	
Time	2.2	0.0001	2.7	0.0001	3.5	0.0001
Time x Manual	5.0	0.0005	5.5	0.0003	8.2	0.0001
Time x Jefferson	7.3	0.0043	7.5	0.0039	7.1	0.0036
Time x Control	0.0		0.0		0.0	
Number of Observations	11069		11069		11064	

FIG. A-15

PFP Effect—High Schools—*ITBS* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	40.7	0.0001	41.0	0.0001	50.8	0.0001
Pilot : Manual	-17.7	0.0056	-12.8	0.0377	-16.2	0.0369
Pilot : Jefferson	-15.1	0.0923	-26.0	0.0038	-26.1	0.0196
Control	0.0		0.0		0.0	
Time	3.5	0.0001	3.8	0.0001	4.2	0.0001
Time x Manual	3.2	0.0171	3.9	0.0042	4.8	0.0002
Time x Jefferson	12.9	0.0001	13.4	0.0001	12.6	0.0001
Time x Control	0.0		0.0		0.0	
Number of Observations	16855		16855		16851	

FIG. A-16

PFP Effect—High Schools—*CSAP* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	50.5	0.0001	52.3	0.0001	65.3	0.0001
Pilot: Manual	-12.3	0.0712	-0.8	0.8485	-4.6	0.3337
Pilot: Jefferson	6.0	0.5708	-6.6	0.1866	-9.8	0.1106
Control	0.0		0.0		0.0	
Time	0.2	0.1902	0.4	0.0454	1.2	0.0001
Time x Manual	0.3	0.7972	-0.2	0.8477	1.0	0.3296
Time x Jefferson	0.2	0.7223	0.5	0.4604	1.5	0.0155
Time x Control	0.0		0.0		0.0	
Number of Observations	20831		20831		20827	

FIG. A-17

PFP Effect—High Schools—*CSAP* Writing NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	50.4	0.0001	51.0	0.0001	64.0	0.0001
Pilot: Manual	-15.6	0.0300	0.5	0.9216	-6.4	0.2396
Pilot: Jefferson	7.8	0.4560	0.3	0.9482	-1.8	0.7665
Control	0.0		0.0		0.0	
Time	0.2	0.2563	0.5	0.0637	1.7	0.0001
Time x Manual	1.3	0.3844	-0.3	0.8459	1.1	0.3815
Time x Jefferson	-0.5	0.5493	-0.8	0.3552	0.2	0.8025
Time x Control	0.0		0.0		0.0	
Number of Observations	16456		16456		16452	

FIG. A-18

PFP Effect—High Schools—*CSAP* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	51.6	0.0001	51.0	0.0001	60.3	0.0001
Pilot: Manual	-14.9	0.0223	0.5	0.9036	-2.3	0.5870
Pilot: Jefferson	5.6	0.5408	2.0	0.6373	0.1	0.9769
Control	0.0		0.0		0.0	
Time	0.1	0.7309	0.3	0.2707	1.3	0.0001
Time x Manual	1.2	0.3934	-0.5	0.7051	0.5	0.7114
Time x Jefferson	0.1	0.9125	-0.8	0.3556	-0.1	0.8641
Time x Control	0.0		0.0		0.0	
Number of Observations	16649		16649		16645	

FIG. A-19

Individual Growth Models, Unadjusted—Elementary, Middle and High Schools

	ITBS Reading		*ITBS* Language		*ITBS* Math		*CSAP* Reading		*CSAP* Writing		*CSAP* Math	
	β	p(β=0)	β	p(β=0)	β	p(β=0)	β	p(β=0)	β	p(β=0)	β	p(β=0)
Elementary Schools												
Intercept	44.5	0.0001	45.0	0.0010	44.9	0.0001	51.5	0.0001				
Pilot	-1.3	0.3525	0.2	0.5748	2.2	0.1480	2.0	0.1823				
Control	0.0		0.0		0.0		0.0					
Time	0.8	0.0001	0.2	0.0147	-0.5	0.0001	1.0	0.0001				
Time x Pilot	-0.4	0.0143	-1.1	0.0001	-0.8	0.0001	-0.5	0.0183				
Time x Control	0.0		0.0		0.0		0.0					
Observations	19749		17837		18248		15417					
Middle Schools												
Intercept	43.7	0.0001	49.1	0.0001	40.9	0.0001	50.9	0.0001	51.7	0.0001	52.9	0.0001
Pilot	-8.9	0.2022	-8.1	0.1596	-5.9	0.3276	-8.5	0.2040	-9.7	0.1444	-14.1	0.0453
Control	0.0		0.0		0.0		0.0		0.0		0.0	
Time	0.1	0.5632	-2.4	0.0001	2.7	0.0001	0.4	0.0007	0.01	0.9581	-1.0	0.0208
Time x Pilot	-0.1	0.8757	-0.1	0.7405	0.5	0.1406	0.004	0.9902	0.9	0.0548	4.9	0.0001
Time x Control	0.0		0.0		0.0		0.0		0.0		0.0	
Observations	15384		14996		14504		13125		9505		6572	
High Schools												
Intercept	48.3	0.0001	44.5	0.0001	48.0	0.0001	53.4	0.0001				
Pilot: Manual	-14.1	0.0279	-15.3	0.0430	-15.0	0.0143	-23.1	0.0053				
Pilot: Jefferson	9.2	0.3372			9.9	0.2474	6.7	0.5142				
Control	0.0		0.0		0.0		0.0					
Time	0.4	0.0319	0.2	0.4981	-0.2	0.5869	-2.2	0.0010				
Time x Manual	0.8	0.3434	4.3	0.0732	2.9	0.0231	2.5	0.0021				
Time x Jefferson	-0.9	0.0830			-2.7	0.0180	-0.03	0.9702				
Time x Control	0.0		0.0		0.0		0.0					
Observations	7229		3530		5105		6213					

FIG. A-20

PFP Effect by Maximum Rubric Level—Elementary Schools—*ITBS* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	9.0	0.0034	8.6	0.0022	17.6	0.0001
Rubric Level 4	12.5	0.0094	11.5	0.0175	11.1	0.0198
Rubric Level 3	6.8	0.0044	6.6	0.0032	7.2	0.0012
Rubric Level 2	6.4	0.0056	6.2	0.0036	6.4	0.0028
Rubric Level 1	0.0		0.0		0.0	
Time	11.8	0.0036	11.4	0.0031	11.7	0.0021
Time Squared	-5.2	0.0126	-5.1	0.0133	-5.3	0.0096
Time x Rubric Level 4	-17.7	0.0039	-16.7	0.0064	-16.2	0.0071
Time x Rubric Level 3	-10.7	0.0100	-10.4	0.0087	-11.0	0.0052
Time x Rubric Level 2	-12.2	0.0029	-11.8	0.0025	-12.0	0.0020
Time x Rubric Level 1	0.0		0.0		0.0	
Time Squared x Rubric Level 4	6.5	0.0061	6.3	0.0079	6.2	0.0078
Time Squared x Rubric Level 3	4.6	0.0283	4.5	0.0290	4.7	0.0209
Time Squared x Rubric Level 2	5.4	0.0107	5.3	0.0115	5.4	0.0087
Time Squared x Rubric Level 1	0.0		0.0		0.0	
Last score	0.7	0.0001	0.7	0.0001	0.6	0.0001
	Least Square Means					
Max Rubric Level 4	51.1		51.2		50.9	
Max Rubric Level 3	49.3		49.6		49.7	
Max Rubric Level 2	49.1		49.5		49.5	
Max Rubric Level 1	42.6		43.1		42.7	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	1.8	0.1711	1.5	0.2475	1.2	0.3681
Difference Rubric Level 4 - Level 2	1.9	0.1497	1.6	0.2275	1.4	0.3106
Difference Rubric Level 4 - Level 1	8.5	0.0134	8.1	0.0169	8.1	0.0151
Difference Rubric Level 3 - Level 2	0.1	0.7440	0.1	0.7982	0.2	0.6376
Difference Rubric Level 3 - Level 1	6.7	0.0357	6.6	0.0360	6.9	0.0243
Difference Rubric Level 2 - Level 1	6.5	0.0388	6.5	0.0384	6.8	0.0278
Number of Observations	8554		8554		8554	

FIG. A-21

PFP Effect by Maximum Rubric Level—Elementary Schools—*ITBS* Language NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	19.8	0.0001	21.8	0.0001	30.6	0.0001
Rubric Level 4	25.3	0.0237	30.0	0.0075	30.1	0.0063
Rubric Level 3	2.4	0.4157	2.3	0.4535	2.6	0.3798
Rubric Level 2	-1.7	0.5448	-1.8	0.5372	-1.7	0.5450
Rubric Level 1	0.0		0.0		0.0	
Time	5.5	0.2663	12.2	0.0186	12.1	0.0166
Time Squared	-4.6	0.0645	-7.1	0.0054	-6.9	0.0065
Time x Rubric Level 4	-30.1	0.0151	-40.1	0.0014	-38.1	0.0020
Time x Rubric Level 3	-15.9	0.0019	-20.3	0.0001	-19.7	0.0002
Time x Rubric Level 2	-4.0	0.4199	-8.3	0.1100	-8.3	0.1010
Time x Rubric Level 1	0.0		0.0		0.0	
Time Squared x Rubric Level 4	9.9	0.0064	13.3	0.0003	12.4	0.0007
Time Squared x Rubric Level 3	7.9	0.0018	9.9	0.0001	9.4	0.0002
Time Squared x Rubric Level 2	3.6	0.1461	5.7	0.0279	5.5	0.0297
Time Squared x Rubric Level 1	0.0		0.0		0.0	
Last score	0.5	0.0001	0.6	0.0001	0.5	0.0001
	Least Square Means					
Max Rubric Level 4	53.6		55.8		56.8	
Max Rubric Level 3	43.2		43.8		44.6	
Max Rubric Level 2	42.1		42.9		43.6	
Max Rubric Level 1	38.3		38.7		39.9	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	10.5	0.0070	12.0	0.0020	12.2	0.0013
Difference Rubric Level 4 - Level 2	11.5	0.0031	12.9	0.0010	13.2	0.0006
Difference Rubric Level 4 - Level 1	15.3	0.0029	17.1	0.0009	16.9	0.0009
Difference Rubric Level 3 - Level 2	1.0	0.0674	0.8	0.1427	1.0	0.0863
Difference Rubric Level 3 - Level 1	4.9	0.1559	5.1	0.1383	4.7	0.1624
Difference Rubric Level 2 - Level 1	3.8	0.2609	4.2	0.2134	3.7	0.2642
Number of Observations	5324		5324		5324	

FIG. A-22

PFP Effect by Maximum Rubric Level—Elementary Schools—*ITBS* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	11.0	0.0031	10.1	0.0053	17.6	0.0001
Rubric Level 4	-18.2	0.0071	-18.4	0.0069	-18.7	0.0049
Rubric Level 3	8.9	0.0022	8.4	0.0045	7.6	0.0087
Rubric Level 2	8.7	0.0018	8.6	0.0021	7.9	0.0041
Rubric Level 1	0.0		0.0		0.0	
Time	18.3	0.0001	16.1	0.0009	15.0	0.0016
Time Squared	-9.2	0.0001	-7.9	0.001	-7.3	0.0020
Time x Rubric Level 4	3.6	0.6483	6.5	0.4127	7.1	0.3627
Time x Rubric Level 3	-18.7	0.0002	-15.3	0.0023	-14.4	0.0037
Time x Rubric Level 2	-20.2	0.0001	-17.4	0.0004	-16.1	0.0008
Time x Rubric Level 1	0.0		0.0		0.0	
Time Squared x Rubric Level 4	4.9	0.0782	3.5	0.2193	2.9	0.2901
Time Squared x Rubric Level 3	9.0	0.0002	7.5	0.0023	6.9	0.0042
Time Squared x Rubric Level 2	9.6	0.0001	8.2	0.0008	7.5	0.0017
Time Squared x Rubric Level 1	0.0		0.0		0.0	
Last score	0.6	0.0001	0.6	0.0001	0.5	0.0001
	Least Square Means					
Max Rubric Level 4	39.4		40.2		39.7	
Max Rubric Level 3	46.8		47.6		47.1	
Max Rubric Level 2	46.4		47.3		47.0	
Max Rubric Level 1	34.4		36.0		36.9	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	-7.4	0.0001	-7.4	0.0001	-7.4	0.0001
Difference Rubric Level 4 - Level 2	-7.0	0.0001	-7.1	0.0001	-7.3	0.0001
Difference Rubric Level 4 - Level 1	5.0	0.2412	4.2	0.3275	2.8	0.5117
Difference Rubric Level 3 - Level 2	0.5	0.3325	0.3	0.5500	0.1	0.8213
Difference Rubric Level 3 - Level 1	12.4	0.0015	11.6	0.0032	10.2	0.0083
Difference Rubric Level 2 - Level 1	12.0	0.0022	11.3	0.0039	10.1	0.0088
Number of Observations	6825		6825		6825	

FIG. A-23

PFP Effect by Maximum Rubric Level—Elementary Schools—*CSAP* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	22.5	0.0001	21.9	0.0001	31.5	0.0001
Rubric Level 4	-16.6	0.0020	-14.7	0.0054	-15.3	0.0028
Rubric Level 3	-7.3	0.0154	-7.0	0.0184	-7.2	0.0122
Rubric Level 2	-9.2	0.0014	-8.8	0.0020	-9.2	0.0009
Rubric Level 1	0.0		0.0		0.0	
Time	-10.4	0.0001	-9.9	0.0002	-10.3	0.0001
Time Squared	0.4	0.1653	0.3	0.1867	0.3	0.2792
Time x Rubric Level 4	19.8	0.0005	17.3	0.0019	18.1	0.0009
Time x Rubric Level 3	6.3	0.0283	5.9	0.0400	6.2	0.0245
Time x Rubric Level 2	8.5	0.0009	8.0	0.0014	8.5	0.0005
Time x Rubric Level 1	0.0		0.0		0.0	
Time Squared x Rubric Level 4	-2.9	0.0187	-2.3	0.0526	-2.5	0.0368
Time Squared x Rubric Level 3	0.6	0.1177	0.7	0.0869	0.7	0.0911
Time Squared x Rubric Level 2	0.0		0.0		0.0	
Time Squared x Rubric Level 1	0.0		0.0		0.0	
Last score	0.8	0.0001	0.8	0.0001	0.7	0.0001
	Least Square Means					
Max Rubric Level 4	54.8		55.3		54.6	
Max Rubric Level 3	55.1		55.3		54.8	
Max Rubric Level 2	54.3		54.6		54.2	
Max Rubric Level 1						
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	-0.3	0.7850	-0.1	0.9256	-0.2	0.8320
Difference Rubric Level 4 - Level 2	0.5	0.5676	0.7	0.4569	0.4	0.6128
Difference Rubric Level 3 - Level 2	0.8	0.0445	0.8	0.0489	0.6	0.0913
Number of Observations	4556		4556		4556	

FIG.A-24

PFP Effect by Maximum Rubric Level—Elementary Schools—*CSAP* Writing NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	56.6	0.0001	54.9	0.0001	70.6	0.0001
Rubric Level 4	-18.4	0.0597	-17.2	0.0792	-10.6	0.2097
Rubric Level 3	-1.7	0.7539	-1.5	0.7853	-0.8	0.8726
Rubric Level 2	-5.9	0.2699	-4.9	0.3557	-4.3	0.3524
Rubric Level 1	0.0		0.0		0.0	
Time	-3.1	0.5122	-2.5	0.5949	-2.2	0.5830
Time Squared	-0.6	0.1747	-0.7	0.0995	-0.4	0.2672
Time x Rubric Level 4	21.3	0.0231	19.7	0.0349	14.5	0.0723
Time x Rubric Level 3	0.2	0.9648	-0.005	0.9992	0.1	0.9765
Time x Rubric Level 2	5.0	0.2655	4.7	0.2987	3.6	0.3491
Time x Rubric Level 1	0.0		0.0		0.0	
Time Squared x Rubric Level 4	-3.9	0.0421	-3.4	0.0695	-3.0	0.0677
Time Squared x Rubric Level 3	1.2	0.0812	1.3	0.0604	0.7	0.2256
Time Squared x Rubric Level 2	0.0		0.0		0.0	
Time Squared x Rubric Level 1	0.0		0.0		0.0	
	Least Square Means					
Max Rubric Level 4	52.2		52.9		52.4	
Max Rubric Level 3	52.1		52.7		52.0	
Max Rubric Level 2	51.6		52.3		51.9	
Max Rubric Level 1						
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	0.1	0.9241	0.2	0.8667	0.4	0.7143
Difference Rubric Level 4 - Level 2	0.6	0.6769	0.6	0.6524	0.5	0.6633
Difference Rubric Level 3 - Level 2	0.5	0.4657	0.4	0.5261	0.1	0.8634
Number of Observations	5597		5597		5597	

FIG. A-25

PFP Effect by Maximum Rubric Level—Elementary Schools—*CSAP* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	56.4	0.0001	53.4	0.0001	65.4	0.0001
Rubric Level 4	-2.4	0.8360	-3.8	0.7500	3.4	0.7418
Rubric Level 3	-5.5	0.4325	-7.2	0.3120	-4.8	0.4369
Rubric Level 2	0.2	0.9210	-0.4	0.8766	1.0	0.6205
Rubric Level 1	0.0		0.0		0.0	
Time	-0.9	0.8739	1.5	0.7797	2.4	0.6143
Time Squared	-0.3	0.8242	-1.0	0.4827	-1.4	0.2727
Time x Rubric Level 4	7.0	0.6079	6.1	0.6511	-0.8	0.9433
Time x Rubric Level 3	5.6	0.5071	6.0	0.4812	4.8	0.5143
Time x Rubric Level 2	0.0		0.0		0.0	
Time x Rubric Level 1	0.0		0.0		0.0	
Time Squared x Rubric Level 4	-1.5	0.6519	-1.0	0.7726	0.7	0.8027
Time Squared x Rubric Level 3	-1.1	0.6018	-0.9	0.6611	-0.7	0.6995
Time Squared x Rubric Level 2	0.0		0.0		0.0	
Time Squared x Rubric Level 1	0.0		0.0		0.0	
	Least Square Means					
Max Rubric Level 4	57.7		58.4		56.5	
Max Rubric Level 3	53.7		54.7		52.9	
Max Rubric Level 2	53.4		54.0		52.5	
Max Rubric Level 1						
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	4.0	0.0212	3.7	0.0326	3.6	0.0157
Difference Rubric Level 4 - Level 1	4.3	0.0222	4.4	0.0182	4.0	0.0118
Difference Rubric Level 2 - Level 1	0.3	0.8002	0.7	0.5322	0.4	0.6500
Number of Observations	2127		2127		2127	

FIG. A-26

PFP Effect by Maximum Rubric Level—Middle Schools—*ITBS* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	16.1	0.1014	13.0	0.1830	16.8	0.9204
Rubric Level 4	-5.4	0.0872	-5.3	0.1044	-4.1	0.2104
Rubric Level 3	-2.5	0.3520	-2.3	0.3844	-2.4	0.3776
Rubric Level 2	-1.4	0.6007	-1.3	0.6326	-1.0	0.7011
Rubric Level 1	0.0		0.0		0.0	
Time	-0.6	0.2964	-1.1	0.1133	-0.8	0.9782
Time x Rubric Level 4	1.8	0.1920	1.8	0.2163	1.2	0.3844
Time x Rubric Level 3	-0.2	0.7988	-0.3	0.7058	-0.1	0.9253
Time x Rubric Level 2	0.0		0.0		0.0	
Time x Rubric Level 1	0.0		0.0		0.0	
Last score	0.6	0.0001	0.6	0.0001	0.6	0.0001
	Least Square Means					
Max Rubric Level 4	33.1		33.2		33.3	
Max Rubric Level 3	33.6		33.6		33.4	
Max Rubric Level 2	35.0		35.0		34.9	
Max Rubric Level 1						
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	-0.5	0.5553	-0.4	0.6377	-0.1	0.9036
Difference Rubric Level 4 - Level 2	-1.8	0.0221	-1.8	0.0226	-1.6	0.0590
Difference Rubric Level 3 - Level 2	-1.4	0.0182	-1.5	0.0141	-1.5	0.0159
Number of Observations	1789		1789		1789	

FIG. A-27

PFP Effect by Maximum Rubric Level—Middle Schools—*ITBS* Language NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	18.6	0.0995	-6.2	0.7948	22.8	0.0155
Rubric Level 4	-1.2	0.7376	-1.6	0.6658	0.4	0.9092
Rubric Level 3	-1.0	0.7421	-0.9	0.7655	0.3	0.9233
Rubric Level 2	1.0	0.7341	1.0	0.7381	1.8	0.5338
Rubric Level 1	0.0		0.0		0.0	
Time	-2.9	0.0001	-0.1	0.9615	3.3	0.0024
Time x Rubric Level 4	1.5	0.3513	1.8	0.2663	1.2	0.4502
Time x Rubric Level 3	0.7	0.4825	0.5	0.6012	0.8	0.3743
Time x Rubric Level 2	0.0		0.0		0.0	
Time x Rubric Level 1	0.0		0.0		0.0	
Last score	0.6	0.0001	0.6	0.0001	0.5	0.0001
	Least Square Means					
Max Rubric Level 4	40.3		40.0		42.2	
Max Rubric Level 3	39.5		39.0		41.6	
Max Rubric Level 2	40.6		40.3		42.1	
Max Rubric Level 1						
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	0.8	0.3586	1.0	0.2894	0.6	0.5085
Difference Rubric Level 4 - Level 2	-0.4	0.7111	-0.3	0.7473	0.1	0.9309
Difference Rubric Level 3 - Level 2	-1.2	0.1039	-1.3	0.0802	-0.5	0.4907
Number of Observations	1433		1433		1433	

FIG. A-28

PFP Effect by Maximum Rubric Level—Middle Schools—*ITBS* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	19.2	0.0857	28.1	0.1027	33.9	0.0163
Rubric Level 4	6.4	0.1322	0.2	0.9574	2.7	0.5423
Rubric Level 3	-4.4	0.0834	-3.8	0.1253	-5.2	0.0396
Rubric Level 2	-2.5	0.3043	-2.0	0.4039	-2.6	0.2866
Rubric Level 1	0.0		0.0		0.0	
Time	0.2	0.8291	-18.9	0.0001	-1.5	0.4251
Time x Rubric Level 4	-4.4	0.0515	1.5	0.5592	-1.9	0.4455
Time x Rubric Level 3	2.6	0.0580	3.4	0.0129	3.9	0.0054
Time x Rubric Level 2	0.0		0.0		0.0	
Time x Rubric Level 1	0.0		0.0		0.0	
Last score	0.5	0.0001	0.5	0.0001	0.5	0.0001
	Least Square Means					
Max Rubric Level 4	39.2		36.3		38.3	
Max Rubric Level 3	35.1		34.3		35.9	
Max Rubric Level 2	34.5		32.8		34.8	
Max Rubric Level 1						
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	4.1	0.0242	1.8	0.3420	2.4	0.1918
Difference Rubric Level 4 - Level 2	4.7	0.0073	3.2	0.0714	3.5	0.0481
Difference Rubric Level 3 - Level 2	0.6	0.4907	1.4	0.1098	1.1	0.2273
Number of Observations	989		989		989	

FIG. A-29

PFP Effect by Maximum Rubric Level—Middle Schools—*CSAP* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	5.9	0.2367	-5.5	0.3119	0.7	0.8655
Rubric Level 4	0.6	0.8153	1.3	0.6088	2.8	0.2535
Rubric Level 3	2.4	0.2217	2.6	0.1966	2.8	0.1474
Rubric Level 2	0.9	0.6580	1.1	0.5859	1.5	0.4435
Rubric Level 1	0.0		0.0		0.0	
Time	2.5	0.0001	-0.2	0.6899	-0.5	0.2928
Time x Rubric Level 4	-0.1	0.8932	-0.5	0.6576	-0.8	0.4799
Time x Rubric Level 3	-1.8	0.0040	-1.7	0.0075	-1.2	0.0523
Time x Rubric Level 2	0.0		0.0		0.0	
Time x Rubric Level 1	0.0		0.0		0.0	
Last score	0.8	0.0001	0.8	0.0001	0.7	0.0001
	Least Square Means					
Max Rubric Level 4	43.6		43.8		44.0	
Max Rubric Level 3	43.5		43.6		43.5	
Max Rubric Level 2	44.1		44.2		43.6	
Max Rubric Level 1						
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	0.1	0.8232	0.1	0.8225	0.5	0.4541
Difference Rubric Level 4 - Level 2	-0.5	0.4551	-0.4	0.5298	0.4	0.5668
Difference Rubric Level 3 - Level 2	-0.6	0.1749	-0.5	0.2371	-0.1	0.8186
Number of Observations	2238		2238		2238	

FIG. A-30

PFP Effect by Maximum Rubric Level—Middle Schools—*CSAP* Writing NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	41.2	0.0700	33.5	0.3486	44.8	0.0001
Rubric Level 4	-6.7	0.1914	-6.8	0.1840	6.2	0.1593
Rubric Level 3	2.2	0.6298	2.3	0.6234	8.1	0.0439
Rubric Level 2	1.0	0.8321	0.9	0.8427	3.9	0.3286
Rubric Level 1	0.0		0.0		0.0	
Time	1.4	0.1184	2.4	0.527	1.1	0.3960
Time x Rubric Level 4	0.9	0.6198	1.0	0.5935	-2.3	0.1423
Time x Rubric Level 3	-2.6	0.0442	-2.7	0.387	-2.5	0.0233
Time x Rubric Level 2	0.0		0.0		0.0	
Time x Rubric Level 1	0.0		0.0		0.0	
	Least Square Means					
Max Rubric Level 4	37.8		37.8		42.8	
Max Rubric Level 3	41.9		41.9		44.4	
Max Rubric Level 2	44.2		44.2		43.7	
Max Rubric Level 1						
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	-4.1	0.0001	-4.0	0.0001	-1.6	0.0622
Difference Rubric Level 4 - Level 1	-6.4	0.0001	-6.4	0.0001	-0.9	0.3082
Difference Rubric Level 3 - Level 2	-2.3	0.0056	-2.4	0.0047	-0.7	0.3389
Number of Observations	2263		2263		2263	

FIG. A-31

PFP Effect by Maximum Rubric Level—Middle Schools—*CSAP* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	42.9	0.0278	43.1	0.0217	55.4	0.0001
Rubric Level 4	3.4	0.2921	0.8	0.8256	5.2	0.1213
Rubric Level 3	-1.7	0.2268	-1.6	0.1937	-0.8	0.5471
Rubric Level 2	0.0		0.0		0.0	
Time	2.3	0.0220	0.8	0.2481	0.02	0.9836
Time x Rubric Level 4	0.3	0.9071	2.3	0.2693	-0.3	0.8883
Time x Rubric Level 3	0.1	0.9272	0.3	0.7464	0.3	0.8306
Time x Rubric Level 2	0.0		0.0		0.0	
	Least Square Means					
Max Rubric Level 4	48.6		47.8		53.1	
Max Rubric Level 3	43.4		43.6		47.6	
Max Rubric Level 2	44.9		44.9		48.1	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	5.2	0.0031	4.2	0.0249	5.4	0.0012
Difference Rubric Level 4 - Level 2	3.7	0.0282	2.9	0.1024	4.9	0.0019
Difference Rubric Level 3 - Level 2	-1.5	0.1038	-1.4	0.1570	-0.5	0.5543
Number of Observations	1693		1693		1693	

FIG. A-32

PFP Effect by Maximum Rubric Level—Manual High School—*ITBS* Reading NCE Score Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	21.6	0.0001	29.3	0.0001
Rubric Level 4	-0.1	0.9819	-1.7	0.7443
Rubric Level 3	-4.9	0.3329	-5.5	0.2755
Rubric Level 2	-4.6	0.3709	-3.9	0.4503
Rubric Level 1	0.0		0.0	
Time	-5.4	0.0038	-5.7	0.0021
Time x Rubric Level 4	1.9	0.5025	2.2	0.4312
Time x Rubric Level 3	4.7	0.0626	4.7	0.0597
Time x Rubric Level 2	0.0		0.0	
Time x Rubric Level 1	0.0		0.0	
Last score	0.6	0.0001	0.5	0.0001
	Least Square Means			
Max Rubric Level 4	40.6		40.2	
Max Rubric Level 3	37.2		37.6	
Max Rubric Level 2	35.2		37.0	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	3.5	0.0104	2.6	0.0536
Difference Rubric Level 4 - Level 2	5.4	0.0002	3.2	0.0277
Difference Rubric Level 3 - Level 2	1.9	0.1241	0.6	0.6341
Number of Observations	675		672	

FIG. A-33

PFP Effect by Maximum Rubric Level—Manual High School—*ITBS* Language NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	20.4	0.0001	23.8	0.0001
Rubric Level 4	-5.0	0.0408	-5.2	0.0347
Rubric Level 3	-3.6	0.0611	-4.2	0.0328
Rubric Level 2	0.0		0.0	
Time	-10.0	0.0001	-9.5	0.0001
Time x Rubric Level 4	9.3	0.0021	8.6	0.0050
Time x Rubric Level 3	9.8	0.0006	9.8	0.0007
Time x Rubric Level 2	0.0		0.0	
Last score	0.6	0.0001	0.6	0.0001
	Least Square Means			
Max Rubric Level 4	38.1		34.2	
Max Rubric Level 3	39.7		36.0	
Max Rubric Level 2	37.3		34.2	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	-1.6	0.3086	-1.7	0.2874
Difference Rubric Level 4 - Level 2	0.7	0.6042	0.1	0.9680
Difference Rubric Level 3 - Level 2	2.4	0.1128	1.8	0.2392
Number of Observations	417		415	

FIG. A-34

PFP Effect by Maximum Rubric Level—Manual High School—*ITBS* Math NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	20.2	0.0001	21.6	0.0001
Rubric Level 4	-6.9	0.1022	-4.1	0.3463
Rubric Level 3	-3.7	0.0370	-3.5	0.0474
Rubric Level 2	0.0		0.0	
Time	-2.9	0.0946	-2.3	0.1874
Time x Rubric Level 4	9.1	0.0531	6.5	0.1806
Time x Rubric Level 3	6.5	0.0128	6.1	0.0201
Time x Rubric Level 2	0.0		0.0	
Last score	0.5	0.0001	0.5	0.0001
	Least Square Means			
Max Rubric Level 4	35.6		37.2	
Max Rubric Level 3	37.3		37.5	
Max Rubric Level 2	37.4		37.7	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	-1.7	0.4312	-0.3	0.8854
Difference Rubric Level 4 - Level 2	-1.8	0.4130	-0.4	0.8373
Difference Rubric Level 3 - Level 2	-0.05	0.9721	-0.1	0.9225
Number of Observations	559		556	

FIG. A-35

PFP Effect by Maximum Rubric Level—Manual High School—*CSAP* Reading NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	8.6	0.0001	10.7	0.0001
Rubric Level 4	3.2	0.0381	2.3	0.1288
Rubric Level 3	1.8	0.1006	1.2	0.2686
Rubric Level 2	0.0		0.0	
Time	3.2	0.0066	3.4	0.0034
Time x Rubric Level 4	-2.0	0.2926	-1.8	0.3387
Time x Rubric Level 3	-2.4	0.1400	-1.9	0.2556
Time x Rubric Level 2	0.0		0.0	
Last score	0.8	0.0001	0.7	0.0001
	Least Square Means			
Max Rubric Level 4	42.7		43.2	
Max Rubric Level 3	41.2		42.1	
Max Rubric Level 2	40.5		41.8	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	1.6	0.1048	1.1	0.2455
Difference Rubric Level 4 - Level 2	2.2	0.0239	1.4	0.1420
Difference Rubric Level 3 - Level 2	0.6	0.4492	0.3	0.7069
Number of Observations	688		685	

FIG. A-36

PFP Effect by Maximum Rubric Level—Manual High School—*CSAP* Writing NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	5.5	0.1216	7.5	0.0661
Rubric Level 4	12.8	0.0961	11.2	0.1391
Rubric Level 3	0.7	0.9023	1.8	0.7469
Rubric Level 2	0.0		0.0	
Time	4.7	0.1845	6.3	0.0700
Time x Rubric Level 4	-9.8	0.2085	-8.7	0.2598
Time x Rubric Level 3	-2.8	0.6286	-3.9	0.5043
Time x Rubric Level 2	0.0		0.0	
Last score	0.7	0.0001	0.6	0.0001
	Least Square Means			
Max Rubric Level 4	42.0		38.6	
Max Rubric Level 3	36.6		33.9	
Max Rubric Level 2	38.6		35.8	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	5.4	0.0002	4.7	0.0009
Difference Rubric Level 4 - Level 2	3.3	0.0049	2.8	0.0166
Difference Rubric Level 3 - Level 2	-2.0	0.0987	-1.9	0.1192
Number of Observations	334		331	

FIG. A-37

PFP Effect by Maximum Rubric Level—Manual High School—*CSAP* Math NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	15.9	0.0001	17.7	0.0001
Rubric Level 4	-7.9	0.0154	-5.2	0.1233
Rubric Level 3	1.2	0.4143	1.6	0.2912
Rubric Level 2	0.0		0.0	
Time	-3.4	0.0097	-2.7	0.0450
Time x Rubric Level 4	7.3	0.0407	4.8	0.1893
Time x Rubric Level 3	1.3	0.5159	0.3	0.8658
Time x Rubric Level 2	0.0		0.0	
Last score	0.7	0.0001	0.7	0.0001
	Least Square Means			
Max Rubric Level 4	37.8		33.8	
Max Rubric Level 3	43.0		37.7	
Max Rubric Level 2	41.0		35.9	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	-5.2	0.0005	-3.9	0.0125
Difference Rubric Level 4 - Level 2	-3.2	0.0282	-2.1	0.1616
Difference Rubric Level 3 - Level 2	2.0	0.0391	1.8	0.0656
Number of Observations	493		491	

FIG. A-38

PFP Effect by Maximum Rubric Level—Jefferson High School—*ITBS* Reading NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	15.3	0.0001	19.4	0.0001
Rubric Level 4	-6.8	0.0093	-6.4	0.0149
Rubric Level 3	-2.6	0.2383	-2.1	0.3285
Rubric Level 2	-5.1	0.0203	-4.8	0.0281
Rubric Level 1	0.0		0.0	
Time	-0.3	0.9174	-0.4	0.8902
Time x Rubric Level 4	3.1	0.3229	3.0	0.3338
Time x Rubric Level 3	-2.1	0.4352	-2.1	0.4452
Time x Rubric Level 2	0.0		0.0	
Time x Rubric Level 1	0.0		0.0	
Last score	0.8	0.0001	0.7	0.0001
	Least Square Means			
Max Rubric Level 4	54.7		55.7	
Max Rubric Level 3	56.3		57.4	
Max Rubric Level 2	54.9		55.7	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	-1.6	0.1160	-1.7	0.0956
Difference Rubric Level 4 - Level 2	-0.2	0.9217	-0.1	0.9685
Difference Rubric Level 3 - Level 2	1.4	0.2908	1.6	0.2352
Number of Observations	1136		1136	

FIG. A-39

PFP Effect by Maximum Rubric Level—Jefferson High School—*ITBS* Math NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	11.1	0.0001	16.2	0.0001
Rubric Level 4	-9.4	0.0787	-5.9	0.2754
Rubric Level 3	0.8	0.9312	4.5	0.6190
Rubric Level 2	0.0		0.0	
Time	2.1	0.2033	1.9	0.2428
Time x Rubric Level 4	8.1	0.1439	5.0	0.3724
Time x Rubric Level 3	-4.4	0.6324	-7.8	0.3929
Time x Rubric Level 2	0.0		0.0	
Last score	0.8	0.0001	0.7	0.0001
	Least Square Means			
Max Rubric Level 4	52.5		54.3	
Max Rubric Level 3	53.9		55.7	
Max Rubric Level 2	56.2		56.7	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	-1.4	0.6722	-1.4	0.6649
Difference Rubric Level 4 - Level 2	-3.7	0.0612	-2.4	0.2298
Difference Rubric Level 3 - Level 2	-2.3	0.4383	-1.0	0.7389
Number of Observations	807		807	

FIG. A-40

PFP Effect by Maximum Rubric Level—Jefferson High School—*CSAP* Reading NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	8.3	0.0001	13.6	0.0001
Rubric Level 4	2.0	0.1580	2.3	0.1050
Rubric Level 3	2.1	0.0136	2.1	0.0171
Rubric Level 2	0.0		0.0	
Time	3.4	0.0834	3.4	0.0772
Time x Rubric Level 4	-2.9	0.2234	-3.2	0.1758
Time x Rubric Level 3	-5.0	0.0203	-4.8	0.0238
Time x Rubric Level 2	0.0		0.0	
Last score	0.8	0.0001	0.7	0.0001
	Least Square Means			
Max Rubric Level 4	59.5		57.8	
Max Rubric Level 3	58.6		56.9	
Max Rubric Level 2	59.0		57.2	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	0.9	0.2727	1.0	0.2126
Difference Rubric Level 4 - Level 2	0.5	0.6561	0.7	0.5698
Difference Rubric Level 3 - Level 2	-0.3	0.7462	-0.3	0.7687
Number of Observations	920		920	

FIG. A-41

PFP Effect by Maximum Rubric Level—Jefferson High School—*CSAP* Writing NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	19.2	0.0001	25.7	0.0001
Rubric Level 4	-8.0	0.4732	-4.0	0.7222
Rubric Level 3	-6.4	0.2608	-5.8	0.3057
Rubric Level 2	0.0		0.0	
Time	-7.3	0.1418	-7.1	0.1556
Time x Rubric Level 4	10.7	0.3505	6.5	0.5653
Time x Rubric Level 3	6.2	0.3124	5.7	0.3522
Time x Rubric Level 2	0.0		0.0	
Last score	0.8	0.0001	0.7	0.0001
	Least Square Means			
Max Rubric Level 4	60.2		58.2	
Max Rubric Level 3	57.5		55.5	
Max Rubric Level 2	57.9		55.8	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Rubric Level 3	2.7	0.0091	2.7	0.0093
Difference Rubric Level 4 - Rubric Level 2	2.3	0.3023	2.4	0.2864
Difference Rubric Level 3 - Rubric Level 2	-0.4	0.8661	-0.3	0.8995
Number of Observations	471		471	

FIG. A-42

PFP Effect by Maximum Rubric Level—Jefferson High School—*CSAP* Math NCE Scores Weighted Least Squares Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	7.3	0.0001	9.9	0.0001
Rubric Level 4	0.1	0.9624	1.2	0.6894
Rubric Level 3	-10.5	0.2133	-5.7	0.5037
Rubric Level 2	0.0		0.0	
Time	-2.2	0.0527	-2.3	0.0460
Time x Rubric Level 4	1.1	0.7092	0.4	0.9104
Time x Rubric Level 3	10.2	0.2320	5.6	0.5137
Time x Rubric Level 2	0.0		0.0	
Last score	0.9	0.0001	0.8	0.0001
	Least Square Means			
Max Rubric Level 4	60.4		60.8	
Max Rubric Level 3	55.4		57.3	
Max Rubric Level 2	59.5		59.4	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Rubric Level 4 - Level 3	4.9	0.1373	3.6	0.2849
Difference Rubric Level 4 - Level 2	0.9	0.5088	1.4	0.2888
Difference Rubric Level 3 - Level 2	-4.1	0.2042	-2.2	0.5066
Number of Observations	706		706	

FIG. A-43

PFP Effect by Total Objectives Met—Elementary Schools—*ITBS* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	11.7	0.0001	11.7	0.0001	20.6	0.0010
Objectives Met 2	4.1	0.0037	3.8	0.0054	4.0	0.0028
Objectives Met 1	2.2	0.1377	2.2	0.1449	1.9	0.1887
Objectives Met 0	0.0		0.0		0.0	
Time	2.2	0.3579	3.0	0.2007	3.8	0.0998
Time Squared	-0.3	0.7579	-0.5	0.5369	-0.7	0.3484
Time x Objectives Met 2	-1.8	0.4667	-2.0	0.3979	-2.6	0.2640
Time x Objectives Met 1	-0.6	0.8184	-0.9	0.7439	-1.0	0.6934
Time x Objectives Met 0	0.0		0.0		0.0	
Time Squared x Objectives Met 2	0.01	0.9887	0.2	0.8532	0.3	0.6863
Time Squared x Objectives Met 1	-0.6	0.5293	-0.5	0.6110	-0.4	0.6759
Time Squared x Objectives Met 0	0.0		0.0		0.0	
Last score	0.7	0.0001	0.7	0.0001	0.6	0.0001
	Least Square Means					
Objectives Met 2	49.1		49.6		49.5	
Objectives Met 1	47.0		47.6		47.4	
Objectives Met 0	47.5		48.1		48.1	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	2.1	0.0001	2.0	0.0001	2.1	0.0001
Difference Met 2 - Met 0	1.6	0.0457	1.5	0.0699	1.4	0.0783
Difference Met 1 - Met 0	-0.4	0.6175	-0.5	0.5701	-0.7	0.4321
Number of Observations	8608		8608		8608	

FIG. A-44

PFP Effect by Total Objectives Met—Elementary Schools—*ITBS* Language NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	14.7	0.0001	14.8	0.0001	23.8	0.0001
Objectives Met 2	4.9	0.0042	4.6	0.0076	4.9	0.0037
Objectives Met 1	1.9	0.2932	1.7	0.3480	1.6	0.3660
Objectives Met 0	0.0		0.0		0.0	
Time	1.8	0.5456	2.5	0.4164	3.2	0.2955
Time Squared	-0.5	0.6353	-0.2	0.8290	-0.4	0.7097
Time x Objectives Met 2	-4.5	0.1426	-3.2	0.2967	-3.6	0.2401
Time x Objectives Met 1	-0.7	0.8254	0.9	0.7886	1.1	0.7251
Time x Objectives Met 0	0.0		0.0		0.0	
Time Squared x Objectives Met 2	1.3	0.2163	0.6	0.5642	0.7	0.5299
Time Squared x Objectives Met 1	0.0	0.9998	-0.8	0.4670	-0.9	0.4049
Time Squared x Objectives Met 0	0.0		0.0		0.0	
Last score	0.6	0.0001	0.6	0.0001	0.5	0.0001
	Least Square Means					
Objectives Met 2	43.4		44.8		45.5	
Objectives Met 1	41.5		43.0		43.6	
Objectives Met 0	40.5		42.6		43.2	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	1.9	0.0011	1.8	0.0022	1.9	0.0012
Difference Met 2 - Met 0	2.9	0.0056	2.2	0.0352	2.2	0.0319
Difference Met 1 - Met 0	1.0	0.3816	0.4	0.7298	0.3	0.7734
Number of Observations	5412		5412		5412	

FIG. A-45

PFP Effect by Total Objectives Met—Elementary Schools—*ITBS* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	17.5	0.0001	16.1	0.0001	23.0	0.0001
Objectives Met 2	2.6	0.1204	2.6	0.1174	2.7	0.0977
Objectives Met 1	-1.8	0.2922	-1.8	0.2931	-2.2	0.2081
Objectives Met 0	0.0		0.0		0.0	
Time	0.9	0.7487	1.0	0.7175	0.6	0.8260
Time Squared	0.1	0.9201	0.3	0.7745	0.3	0.7527
Time x Objectives Met 2	-3.0	0.2998	-2.1	0.4686	-1.7	0.5384
Time x Objectives Met 1	2.8	0.3703	3.4	0.2789	4.5	0.1391
Time x Objectives Met 0	0.0		0.0		0.0	
Time Squared x Objectives Met 2	0.4	0.6536	0.03	0.9750	-0.02	0.9817
Time Squared x Objectives Met 1	-1.8	0.104	-2.0	0.0700	-2.3	0.0353
Time Squared x Objectives Met 0	0.0		0.0		0.0	
Last score	0.6	0.0001	0.6	0.0001	0.5	0.0001
	Least Square Means					
Objectives Met 2	46.6		47.4		47.0	
Objectives Met 1	43.1		44.1		43.7	
Objectives Met 0	47.0		47.8		47.0	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	3.5	0.0001	3.3	0.0001	3.3	0.0001
Difference Met 2 - Met 0	-0.4	0.6854	-0.5	0.6352	-0.04	0.9670
Difference Met 1 - Met 0	-3.9	0.0003	-3.8	0.0006	-3.3	0.0022
Number of Observations	6870		6870		6870	

FIG. A-46

PFP Effect by Total Objectives Met—Elementary Schools—*CSAP* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	13.6	0.0001	14.0	0.0001	23.2	0.0001
Objectives Met 2	0.6	0.6992	0.5	0.7343	0.5	0.7067
Objectives Met 1	-1.0	0.6027	-1.0	0.6074	-1.1	0.5546
Objectives Met 0	0.0		0.0		0.0	
Time	0.1	0.9573	0.5	0.8028	0.7	0.7433
Time Squared	-1.1	0.1395	-1.2	0.0988	-1.2	0.0964
Time x Objectives Met 2	-2.8	0.2192	-3.0	0.1771	-2.9	0.1869
Time x Objectives Met 1	-1.8	0.5161	-2.0	0.4727	-1.7	0.5198
Time x Objectives Met 0	0.0		0.0		0.0	
Time Squared x Objectives Met 2	1.7	0.0195	1.8	0.0138	1.7	0.0212
Time Squared x Objectives Met 1	1.2	0.168	1.2	0.1528	1.1	0.2170
Time Squared x Objectives Met 0	0.0		0.0		0.0	
Last score	0.8	0.0001	0.8	0.0001	0.7	0.0001
	Least Square Means					
Objectives Met 2	54.9		55.2		54.8	
Objectives Met 1	53.0		53.3		52.7	
Objectives Met 0	52.3		52.7		52.6	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	2.0	0.0001	1.9	0.0001	2.1	0.0001
Difference Met 2 - Met 0	2.7	0.0003	2.5	0.0007	2.2	0.0028
Difference Met 1 - Met 0	0.7	0.3935	0.6	0.4860	0.1	0.8954
Number of Observations	4556		4556		4556	

FIG. A-47

PFP Effect by Total Objectives Met—Elementary Schools—*CSAP* Writing NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	46.4	0.0001	45.8	0.0001	62.4	0.0001
Objectives Met 2	5.8	0.0169	5.7	0.0196	5.8	0.0069
Objectives Met 1	7.6	0.0181	7.4	0.0214	5.4	0.0537
Objectives Met 0	0.0		0.0		0.0	
Time	12.1	0.0007	12.2	0.0007	11.0	0.0004
Time Squared	-4.0	0.0005	-4.0	0.0005	-3.3	0.0012
Time x Objectives Met 2	-12.4	0.0008	-12.5	0.0007	-10.6	0.0009
Time x Objectives Met 1	-10.9	0.0145	-10.5	0.0183	-7.1	0.0687
Time x Objectives Met 0	0.0		0.0		0.0	
Time Squared x Objectives Met 2	4.1	0.0006	4.2	0.0005	3.1	0.0027
Time Squared x Objectives Met 1	3.2	0.0215	3.0	0.0308	1.6	0.1788
Time Squared x Objectives Met 0	0.0		0.0		0.0	
	Least Square Means					
Objectives Met 2	52.0		52.5		52.1	
Objectives Met 1	52.5		52.8		51.5	
Objectives Met 0	50.6		51.3		52.0	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	-0.5	0.5678	-0.3	0.7692	0.5	0.4719
Difference Met 2 - Met 0	1.4	0.2948	1.3	0.3334	0.1	0.9282
Difference Met 1 - Met 0	1.9	0.2039	1.5	0.3039	-0.4	0.7314
Number of Observations	5609		5609		5609	

FIG. A-48

PFP Effect by Total Objectives Met—Elementary Schools—*CSAP* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	69.6	0.0059	62.0	0.0115	74.5	0.0015
Objectives Met 2	-16.9	0.4110	-12.8	0.5423	-12.6	0.4890
Objectives Met 1	-10.1	0.6414	-7.4	0.7382	-9.2	0.6303
Objectives Met 0	0.0		0.0		0.0	
Time	-11.6	0.5786	-6.2	0.7693	-5.1	0.7835
Time Squared	0.1	0.9835	-1.3	0.8154	-1.5	0.7415
Time x Objectives Met 2	14.9	0.485	10.4	0.6343	10.2	0.5879
Time x Objectives Met 1	6.9	0.7615	4.8	0.8347	6.6	0.7416
Time x Objectives Met 0	0.0		0.0		0.0	
Time Squared x Objectives Met 2	-1.0	0.8575	0.3	0.9554	0.1	0.9767
Time Squared x Objectives Met 1	0.1	0.9796	0.6	0.9248	0.1	0.9805
Time Squared x Objectives Met 0	0.0		0.0		0.0	
	Least Square Means					
Objectives Met 2	55.2		55.6		54.0	
Objectives Met 1	51.3		51.2		50.1	
Objectives Met 0	46.9		46.4		45.5	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	3.9	0.0004	4.4	0.0001	3.9	0.0001
Difference Met 2 - Met 0	8.3	0.0088	9.3	0.0042	8.4	0.0028
Difference Met 1 - Met 0	4.4	0.1835	4.8	0.1511	4.6	0.1195
Number of Observations	2117		2117		2117	

A-49

PFP Effect by Total Objectives Met—Middle Schools—*ITBS* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	12.2	0.1460	11.9	0.1495	15.7	0.1295
Objectives Met 2	0.9	0.7559	1.0	0.7190	0.7	0.8124
Objectives Met 1	0.4	0.9128	0.7	0.8383	0.1	0.9643
Objectives Met 0	0.0		0.0		0.0	
Time	0.8	0.7162	0.8	0.7255	0.4	0.8646
Time x Objectives Met 2	-1.6	0.4894	-1.8	0.4261	-1.4	0.5362
Time x Objectives Met 1	-2.4	0.3516	-2.8	0.2704	-2.4	0.3386
Time x Objectives Met 0	0.0		0.0		0.0	
Last score	0.6	0.0001	0.6	0.0001	0.6	0.0001
	Least Square Means					
Objectives Met 2	33.7		33.7		33.9	
Objectives Met 1	32.2		32.1		32.1	
Objectives Met 0	34.7		34.9		35.0	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	1.5	0.0411	1.6	0.0309	1.8	0.0148
Difference Met 2 - Met 0	-1.1	0.2892	-1.2	0.2253	-1.1	0.2885
Difference Met 1 - Met 0	-2.6	0.0271	-2.8	0.0164	-2.9	0.0142
Number of Observations	1800		1800		1800	

FIG. A-50

PFP Effect by Total Objectives Met—Middle Schools—*ITBS* Language NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	11.7	0.1978	11.9	0.1967	21.2	0.1253
Objectives Met 2	6.1	0.0975	6.0	0.1064	2.8	0.4473
Objectives Met 1	3.8	0.3451	3.6	0.3811	0.1	0.9716
Objectives Met 0	0.0		0.0		0.0	
Time	-1.2	0.6548	-1.3	0.6427	-3.3	0.2267
Time x Objectives Met 2	-1.7	0.5422	-1.6	0.5754	0.3	0.9059
Time x Objectives Met 1	1.1	0.7308	1.3	0.6789	3.1	0.3226
Time x Objectives Met 0	0.0		0.0		0.0	
Last score	0.6	0.0001	0.6	0.0001	0.5	0.0001
	Least Square Means					
Objectives Met 2	39.2		39.1		40.7	
Objectives Met 1	40.2		40.2		41.4	
Objectives Met 0	35.1		35.1		37.6	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	-1.1	0.1995	-1.1	0.1921	-0.7	0.3945
Difference Met 2 - Met 0	4.0	0.0020	4.1	0.0019	3.2	0.0139
Difference Met 1 - Met 0	5.1	0.0005	5.2	0.0004	3.9	0.0075
Number of Observations	1453		1453		1453	

FIG. A-51

PFP Effect by Total Objectives Met—Middle Schools—*ITBS* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	8.6	0.4783	18.3	0.2720	24.1	0.2142
Objectives Met 2	5.4	0.5051	-6.7	0.4157	-7.2	0.3771
Objectives Met 1	10.6	0.1966	-2.8	0.7433	-4.0	0.6324
Objectives Met 0	0.0		0.0		0.0	
Time	2.0	0.6750	-6.0	0.2202	-6.2	0.1988
Time x Objectives Met 2	-0.5	0.9225	6.4	0.1928	6.4	0.1851
Time x Objectives Met 1	-4.4	0.3645	3.2	0.5185	3.9	0.4284
Time x Objectives Met 0	0.0		0.0		0.0	
Last score	0.5	0.0001	0.5	0.0001	0.5	0.0001
	Least Square Means					
Objectives Met 2	34.5		34.3		35.0	
Objectives Met 1	34.4		34.1		35.0	
Objectives Met 0	29.7		32.5		33.7	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	0.1	0.9228	0.2	0.8318	0.1	0.9494
Difference Met 2 - Met 0	4.7	0.0813	1.8	0.5160	1.3	0.6278
Difference Met 1 - Met 0	4.6	0.1043	1.6	0.5856	1.3	0.6590
Number of Observations	1011		1011		1011	

FIG. A-52

PFP Effect by Total Objectives Met—Middle Schools—*CSAP* Reading NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	8.1	0.1902	7.8	0.1942	14.6	0.1153
Objectives Met 2	-0.4	0.8442	-0.3	0.8789	-1.0	0.6564
Objectives Met 1	-4.8	0.0490	-4.5	0.0661	-4.6	0.0573
Objectives Met 0	0.0		0.0		0.0	
Time	0.7	0.6842	0.7	0.6866	0.1	0.9297
Time x Objectives Met 2	0.6	0.7087	0.5	0.7952	1.0	0.5754
Time x Objectives Met 1	4.2	0.0284	3.8	0.0502	3.9	0.0427
Time x Objectives Met 0	0.0		0.0		0.0	
Last score	0.8	0.0001	0.8	0.0001	0.7	0.0001
	Least Square Means					
Objectives Met 2	43.6		43.6		43.4	
Objectives Met 1	43.6		43.5		43.3	
Objectives Met 0	43.2		43.4		43.2	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	0.004	0.9948	0.1	0.9155	0.1	0.9184
Difference Met 2 - Met 0	0.4	0.6358	0.2	0.7761	0.2	0.7711
Difference Met 1 - Met 0	0.4	0.6820	0.2	0.8542	0.2	0.8479
Number of Observations	2223		2223		2223	

FIG. A-53

PFP Effect by Total Objectives Met—Middle Schools—*CSAP* Writing NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	15.4	0.1329	15.4	0.1329	27.3	0.0811
Objectives Met 2	-2.4	0.4525	-2.4	0.4550	-1.8	0.5649
Objectives Met 1	5.3	0.2214	5.2	0.2253	4.8	0.2466
Objectives Met 0	0.0		0.0		0.0	
Time	0.6	0.7880	0.6	0.7844	0.2	0.9309
Time x Objectives Met 2	1.0	0.6626	1.0	0.6576	1.2	0.5994
Time x Objectives Met 1	-3.4	0.2419	-3.3	0.2553	-2.6	0.3435
Time x Objectives Met 0	0.0		0.0		0.0	
Last score	0.7	0.0001	0.7	0.0001	0.6	0.0001
	Least Square Means					
Objectives Met 2	43.9		43.9		45.1	
Objectives Met 1	44.7		44.7		45.6	
Objectives Met 0	44.7		44.7		45.1	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	-0.7	0.4460	-0.7	0.4398	-0.5	0.5755
Difference Met 2 - Met 0	-0.8	0.5570	-0.8	0.5862	0.1	0.9483
Difference Met 1 - Met 0	-0.1	0.9626	-0.02	0.9912	0.6	0.7002
Number of Observations	1325		1325		1325	

FIG. A-54

PFP Effect by Total Objectives Met—Middle Schools—*CSAP* Math NCE Scores Weighted Two-Stage Hierarchical Linear Model

	Unadjusted		Adjusted for School Factors		Adjusted for School and Student Factors	
	β	p(β=0)	β	p(β=0)	β	p(β=0)
Intercept	2.9	0.7182	4.0	0.6373	8.8	0.3992
Objectives Met 2	0.9	0.8783	0.3	0.9666	1.1	0.8542
Objectives Met 1	-13.6	0.1565	-14.7	0.1294	-15.4	0.1087
Objectives Met 0	0.0		0.0		0.0	
Time	4.6	0.1641	3.5	0.3126	3.2	0.3511
Time x Objectives Met 2	0.5	0.8903	0.9	0.7872	0.9	0.7931
Time x Objectives Met 1	8.1	0.1068	8.9	0.0832	10.1	0.0462
Time x Objectives Met 0	0.0		0.0		0.0	
Last score	0.7	0.0001	0.7	0.0001	0.7	0.0001
	Least Square Means					
Objectives Met 2	45.9		45.9		46.8	
Objectives Met 1	44.8		44.9		46.5	
Objectives Met 0	44.1		44.0		44.1	
	Difference	p (difference>0)	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	1.1	0.3969	1.0	0.4362	0.3	0.7901
Difference Met 2 - Met 0	1.8	0.2078	1.9	0.1809	2.7	0.0573
Difference Met 1 - Met 0	0.7	0.7198	0.9	0.6408	2.3	0.2109
Number of Observations	950		950		950	

FIG. A-55

PFP Effect by Total Objectives Met—Manual High School—*ITBS* Reading NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	9.1	0.0004	15.4	0.0001
Objectives Met 2	7.6	0.0019	7.8	0.0012
Objectives Met 1	2.2	0.4595	1.5	0.5975
Objectives Met 0	0.0		0.0	
Time	5.4	0.0991	4.4	0.1681
Time x Objectives Met 2	-9.5	0.0064	-8.7	0.1140
Time x Objectives Met 1	-3.4	0.4448	-3.0	0.5051
Time x Objectives Met 0	0.0		0.0	
Last score	0.6	0.0001	0.5	0.0001
	Least Square Means			
Objectives Met 2	36.8		37.0	
Objectives Met 1	34.2		33.3	
Objectives Met 0	33.6		33.1	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	2.6	0.1073	3.7	0.0208
Difference Met 2 - Met 0	3.2	0.0697	3.8	0.0259
Difference Met 1 - Met 0	0.6	0.7885	0.2	0.9352
Number of Observations	692		689	

FIG. A-56

PFP Effect by Total Objectives Met—Manual High School—*ITBS* Language NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	19.7	0.0015	23.4	0.0013
Objectives Met 2	-4.9	0.4250	-3.1	0.6110
Objectives Met 1	3.7	0.3393	4.0	0.2983
Objectives Met 0	0.0		0.0	
Time	-8.4	0.1331	-7.4	0.1877
Time x Objectives Met 2	5.0	0.3798	3.9	0.4923
Time x Objectives Met 1	0.0		0.0	
Time x Objectives Met 0	0.0		0.0	
Last score	0.7	0.0001	0.6	0.0001
	Least Square Means			
Objectives Met 2	36.6		32.2	
Objectives Met 1	42.1		37.0	
Objectives Met 0				
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	-5.5	0.0415	-4.7	0.0782
Difference Met 2 - Met 0				
Difference Met 1 - Met 0				
Number of Observations	430		428	

FIG. A-57

PFP Effect by Total Objectives Met—Manual High School—*ITBS* Math NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	17.5	0.0001	20.5	0.0001
Objectives Met 2	0.1	0.9438	0.5	0.7947
Objectives Met 1	7.7	0.0046	7.0	0.0106
Objectives Met 0	0.0		0.0	
Time	-2.3	0.2516	-1.8	0.3709
Time x Objectives Met 2	2.8	0.2599	2.7	0.2709
Time x Objectives Met 1	0.0		0.0	
Time x Objectives Met 0	0.0		0.0	
Last score	0.5	0.0001	0.5	0.0001
	Least Square Means			
Objectives Met 2	37.1		38.0	
Objectives Met 1				
Objectives Met 0	35.4		35.9	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1				
Difference Met 2 - Met 0	1.7	0.1693	2.0	0.1163
Difference Met 1 - Met 0				
Number of Observations	588		585	

FIG. A-58

PFP Effect by Total Objectives Met—Manual High School—*CSAP* Reading NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	9.8	0.0001	12.2	0.0001
Objectives Met 2	0.4	0.7669	-0.2	0.8870
Objectives Met 1	0.3	0.8503	-0.5	0.7879
Objectives Met 0	0.0		0.0	
Time	2.8	0.1388	2.2	0.2353
Time x Objectives Met 2	-1.3	0.5114	-0.1	0.9692
Time x Objectives Met 1	-0.3	0.9107	1.2	0.6641
Time x Objectives Met 0	0.0		0.0	
Last score	0.8	0.0001	0.7	0.0001
	Least Square Means			
Objectives Met 2	41.4		42.5	
Objectives Met 1	41.8		42.9	
Objectives Met 0	41.6		42.7	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	-0.4	0.7202	-0.4	0.7281
Difference Met 2 - Met 0	-0.2	0.8278	-0.2	0.8131
Difference Met 1 - Met 0	0.2	0.9038	0.1	0.9232
Number of Observations	690		687	

FIG. A-59

PFP Effect by Total Objectives Met—Manual High School—*CSAP* Writing NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	13.7	0.0577	13.9	0.0563
Objectives Met 2	-8.0	0.2969	-4.0	0.5972
Objectives Met 1	0.02	0.9940	-0.3	0.8834
Objectives Met 0	0.0		0.0	
Time	-5.1	0.4615	-0.4	0.9578
Time x Objectives Met 2	9.6	0.2031	6.0	0.4195
Time x Objectives Met 1	0.0		0.0	
Time x Objectives Met 0	0.0		0.0	
Last score	0.7	0.0001	0.6	0.0001
	Least Square Means			
Objectives Met 2	39.4		36.8	
Objectives Met 1	38.1		34.7	
Objectives Met 0				
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	1.3	0.4848	2.1	0.2540
Difference Met 2 - Met 0				
Difference Met 1 - Met 0				
Number of Observations	336		333	

FIG. A-60

PFP Effect by Total Objectives Met—Manual High School—*CSAP* Math NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	15.7	0.0001	17.9	0.0001
Objectives Met 2	0.4	0.7901	1.2	0.4171
Objectives Met 1	1.1	0.6658	1.0	0.6789
Objectives Met 0	0.0		0.0	
Time	-3.0	0.0240	-2.8	0.0417
Time x Objectives Met 2	0.9	0.6240	1.0	0.5906
Time x Objectives Met 1	0.0		0.0	
Time x Objectives Met 0	0.0		0.0	
Last score	0.7	0.0001	0.6	0.0001
	Least Square Means			
Objectives Met 2	42.1		37.4	
Objectives Met 1				
Objectives Met 0	41.1		35.6	
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1				
Difference Met 2 - Met 0	1.0	0.2778	1.8	0.0585
Difference Met 1 - Met 0				
Number of Observations	512		510	

FIG. A-61

PFP Effect by Total Objectives Met—Jefferson High School—*ITBS* Reading NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	5.9	0.0032	10.3	0.0001
Objectives Met 2	5.4	0.0033	5.4	0.0027
Objectives Met 1	0.0		0.0	
Objectives Met 0				
Time	-1.2	0.1007	-1.2	0.0820
Last score	0.8	0.0001	0.8	0.0001
	Least Square Means			
Objectives Met 2	56.0		57.1	
Objectives Met 1	50.6		51.7	
Objectives Met 0				
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	5.4	0.0033	5.4	0.0027
Difference Met 2 - Met 0				
Difference Met 1 - Met 0				
Number of Observations	1137		1137	

FIG. A-62

PFP Effect by Total Objectives Met—Jefferson High School—*ITBS* Math NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	8.4	0.0001	13.5	0.0001
Objectives Met 2	1.4	0.4533	1.6	0.3701
Objectives Met 1	0.0		0.0	
Objectives Met 0				
Time	-0.04	0.9754	-0.1	0.9248
Last score	0.8	0.0001	0.7	0.0001
	Least Square Means			
Objectives Met 2	54.6		55.7	
Objectives Met 1	53.2		54.1	
Objectives Met 0				
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	1.4	0.4533	1.6	0.3701
Difference Met 2 - Met 0				
Difference Met 1 - Met 0				
Number of Observations	809		809	

FIG. A-63

PFP Effect by Total Objectives Met—Jefferson High School—*CSAP* Reading NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	7.7	0.0001	12.8	0.0001
Objectives Met 2	1.1	0.4206	1.4	0.2837
Objectives Met 1	0.0		0.0	
Objectives Met 0				
Time	0.3	0.6477	0.3	0.6029
Last score	0.8	0.0001	0.8	0.0001
	Least Square Means			
Objectives Met 2	58.9		57.0	
Objectives Met 1	57.8		55.6	
Objectives Met 0				
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	1.1	0.4206	1.4	0.2837
Difference Met 2 - Met 0				
Difference Met 1 - Met 0				
Number of Observations	917		917	

FIG. A-64

PFP Effect by Total Objectives Met—Jefferson High School—*CSAP* Math NCE Scores Weighted Least Squares Linear Regression Model

	Unadjusted		Adjusted for Student Factors	
	β	p(β=0)	β	p(β=0)
Intercept	6.2	0.0001	8.8	0.0001
Objectives Met 2	1.6	0.1486	1.7	0.1249
Objectives Met 1	0.0		0.0	
Objectives Met 0				
Time	-2.1	0.0039	-2.1	0.0049
Last score	0.9	0.0001	0.8	0.0001
	Least Square Means			
Objectives Met 2	60.1		60.2	
Objectives Met 1	58.5		58.5	
Objectives Met 0				
	Difference	p (difference>0)	Difference	p (difference>0)
Difference Met 2 - Met 1	1.6	0.1486	1.7	0.1249
Difference Met 2 - Met 0				
Difference Met 1 - Met 0				
Number of Observations	704		704	

Notes

COMMUNITY TRAINING
AND ASSISTANCE CENTER

30 WINTER STREET • BOSTON, MA 02108
TEL: 617.423.1444 • E-MAIL: ctac@ctacusa.com
www.ctacusa.com