BRINGING URBAN HIGH SCHOOL REFORM TO SCALE: RAPIDLY MOVING DRAMATIC NUMBERS OF STUDENTS TO PROFICIENT PERFORMANCE

By

Joseph Christopher Burks, Jr.
B.S., Centre College, 1975
M.Ed., University of Louisville, 1978

Glenn Stephen Baete
B.A., University of Louisville, 1991
M.Ed., University of Louisville, 1996

Martin Anthony Pollio
B.A., Indiana University, 1993
M.Ed., Eastern Kentucky University, 1995

A Dissertation
Submitted to the Faculty of the
College of Education and Human Development of the University of Louisville
In Partial Fulfillment of the Requirements
for the Degree of

Doctor of Education

Department of Leadership, Foundations & Human Resource Education
University of Louisville
Louisville, Kentucky

May 2012

BRINGING URBAN HIGH SCHOOL REFORM TO SCALE: RAPIDLY MOVING DRAMATIC NUMBERS OF STUDENTS TO PROFICIENT PERFORMANCE

By

Joseph Christopher Burks, Jr.
B.S., Centre College, 1975
M.Ed., University of Louisville, 1978

Glenn Stephen Baete
B.A., University of Louisville, 1991
M.Ed., University of Louisville, 1996

Martin Anthony Pollio
B.A., Indiana University, 1993
M.Ed., Eastern Kentucky University, 1995

A Dissertation Approved on

April 4, 2012

by the following Dissertation Committee:

_____
Dissertation Director

_____

_____

_____

DEDICATION

This dissertation is dedicated to the teachers and leaders of Jefferson County Public Schools featured in this study who risked more than others thought was safe and expected more from their students than others thought was possible.

ACKNOWLEDGEMENTS

Joe Burks-- My sincere thanks and appreciation to Dr. Craig Hochbein, Dr. Marco Muñoz, Dr. Sam Stringfield, Dr. Diana Oxley, Dr. Molly Sullivan, and Dr. Joe Petrosko for their instruction, guidance, and assistance critical to this capstone. Their expertise, passion, and encouragement reminded me to persist with purpose. I also want to thank Dr. Robert Rodosky and Dr. Brian Shumate who initially challenged and influenced me to pursue this journey. Thanks to my co-researchers, Glenn Baete and Marty Pollio, whose energy and drive helped me endure the classes, comprehensive exams, and capstone marathon. Finally, I thank God for this opportunity to serve Him and for blessing me with an unbelievably supportive family. Thanks to my parents, Joe and Kathleen Burks, for teaching me values, work ethic, and perseverance; to my children, Cindy and Adam, and my grandchildren, Bailey, Jedidiah, Micah, and Elijah, for providing me the inspiration to set a family standard; and to my wife, Joan, who always sustains me with unfailing and undeserved support, sacrifice, and love.

Glenn Baete— I would like to thank Dr. Marco Muñoz for his support, direction, and confidence in me during my doctoral studies. His mentoring fueled my desire to learn. I am also indebted to Dr. Craig Hochbein, Dr. Diana Oxley, and Dr. Sam Stringfield for their guidance on this project. They caused me to stretch my thinking to levels I never thought was possible. In addition, I would like to thank Joe Burks and Marty Pollio for their unwavering dedication to the students of Jefferson County, their devotion to our coursework, and their collaboration on this capstone. I owe my parents, Chester and Rita Baete, a tremendous debt of gratitude for teaching me to place family first, to love learning, and to pursue goals with passion. Finally, this work would not have been possible without the support of fantastic family and friends; my children, Emily and Andrew; and most importantly, my wife, Carolyn. She has been supportive in my pursuits and steadfast in her sacrifice and love.

Marty Pollio-- I would like to thank Dr. Craig Hochbein for his guidance, assistance, and advocacy through the difficult process of completing this capstone. Dr. Hochbein's teaching abilities and leadership have caused me to grow as both a researcher and practitioner over the past three years. I would also like to thank Dr. Sam Stringfield, Dr. Marco Muñoz, and Dr. Diana Oxley for their direction through this enormous challenge. Also, I would not have made it to this point without the drive and determination of my colleagues, Glenn Baete and Joe Burks. Thanks to my parents, Mike and Ann Pollio, for teaching me the importance of life-long education and for my Mom's assistance in the writing process. Finally and most importantly, I would like to thank my wife, Jessica and my daughter, Genna. Over the past three years, Jessica has not only had to endure the life of a principal's wife but also that of a doctoral student’s wife. She has been patient, kind, supportive, and understanding throughout this long process. I will never forget her support and love.

ABSTRACT
BRINGING URBAN HIGH SCHOOL REFORM TO SCALE: RAPIDLY MOVING DRAMATIC NUMBERS OF STUDENTS TO PROFICIENT PERFORMANCE

Joseph C. Burks, Jr.
Glenn S. Baete
Martin A. Pollio

April 4, 2012

This capstone is an examination of the effects of Project Proficiency (PP), a high school scale-up effort implemented in a large urban school district. The capstone establishes recurring issues with high school reform, identifies problems with taking reform efforts to scale, examines the importance of instructional coherence, and provides an overview of PP. The next part of the capstone contains three studies that examine the effects of PP on grading practices, variance in relationships within and between PP classrooms, and the reform's success with academic performance of the district's most at-risk students. The final part of the capstone contains an executive summary that compiles the findings and discusses how PP contains elements of a scalable reform effort.

In three separate studies, researchers used nonequivalent control group designs to compare mathematics achievement results in 11 high schools one year prior to and one year following PP implementation. In one study, researchers employed regression analysis to measure the association between classroom grades and student achievement. The study argued that the use of standards-based grading in PP schools increased the association between grades and test scores. In the second study, researchers use

hierarchical linear modeling (HLM) to control for individual and classroom socioeconomic status (SES) and prior student achievement to determine if changes in instructional practices yielded academic gains. The HLM analysis indicated that statistically significant gains existed in mathematics achievement and significant reductions in between-classroom variance.

In the third study, researchers examined PP's impact on performance with students who met dropout predictive criteria established by Balfanz, Herzog, and MacIver (2007). The results suggest that PP students most at-risk for dropout realized statistically significant achievement gains from grades 8 to 11. The capstone concludes with an executive summary that articulates PP's impact on student achievement and suggests that PP is a viable and scalable urban high school reform based on its elements of depth, spread, shift of ownership, and sustainability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

URBAN HIGH SCHOOL REFORM TO SCALE: RAPIDLY MOVING DRAMATIC NUMBERS OF STUDENTS TO PROFICIENT PERFORMANCE

High School Reform

In 1932, a group of over 300 colleges and universities partnered with 30 high schools in one the 21st century's first high school reform projects. In what was later called the 8-Year Study, a cross-section of American high schools had the freedom to redesign their curriculum without the fear of graduates denied admittance to college for lacking traditional entrance requirements (Aiken, 1942). Given the charge to redefine the purpose of their high schools, 8-Year participants connected teaching and learning to emerging knowledge of human growth and development, experimented with longer class periods, eliminated divisions between curricular and extracurricular activities, and modified graduation requirements. In turn, the participating colleges unconditionally admitted project high school graduates who secured the principal's recommendation and submitted a record of their involvement in activities, interests, and academic work. The study revealed that students from 8-Year Schools were neither ill-prepared for college work, nor displayed negative differences in college performance than their non-8 Year counterparts (Aiken, 1942). Subsequent analysis revealed that secondary schools with the most progressive reform strategies produced gains that exceeded non-participating schools.

Seven decades after Aiken (1942) reported the findings of the 8-Year Study, educators continue to grapple with high school reform. The desire for educators to make significant and lasting instructional improvements is of importance as significant and lasting instructional improvements is of importance as public confidence in public schools is at record-low levels (Gallup, 2011). Modern policymakers seek to create a 21st century workforce with globally competitive skills, improve American productivity and economic growth, and continue the US role as world power (National Center on Education and the Economy, 2008). National leaders have asserted that education, particularly at the high school level, will stem the tide of mediocrity that threatens America's prosperity (United States Department of Education, 1983).

Since the work of Aiken (1942), a plethora of studies over the past half-century have investigated the efforts taken by districts and high schools to implement large-scale reforms (e.g. Berends, Bodilly, & Kirby, 2002; Bryk, 2010; Consortium on Chicago School Research, 2010; Crandall et al., 1982; Cuban, 1984; Darling-Hammond, 2006; Datnow, 2005; Datnow & Stringfield, 2000; Fullan, 2000; Fullan, Bertani, & Quinn, 2004; Fullan & Pomfret, 1977; Holdzkom, 2002; McLaughlin, 1990; Oxley & Kassissieh, 2008; Oxley & Luers, 2010; Rorrer, Skrla, & Schuerich, 2008; Quint, 2006; Stringfield, Reynolds, & Schaffer, 2008). These researchers identified successes and challenges with implementing high school reform and cited key considerations including identifying the purpose of the reform, creating structures necessary for successful implementation, and providing effective internal and external supports to scale up those reforms.

Purpose of reform. For the past twenty years, the desire by school districts to meet state and federal accountability measures and avoid sanctions drives American high school reform (Datnow, 2005; Fullan, 2000, 2011; No Child Left Behind, 2001; Race to the Top, 2010). While these mandates do not require a school to identify a specific reform initiative, three of the four current federal turnaround models require high schools to redesign structures and use frequent benchmarking to measure progress (School Improvement Fund, 2010). Co-existent with meeting accountability requirements, many reform efforts sought to assist students who enter high schools ill-prepared for the rigor necessary to succeed (Datnow, 2005; Holdzkom, 2002; Quint, 2006). Thirty years before Race to the Top (RTTT), Fullan and Pomfret (1977) discussed the moral imperative for schools to raise student achievement levels and close achievement gaps for all individuals and schools. Federal legislation supports this trend in support of individual student accountability as No Child Left Behind (NCLB) guidelines require schools to show achievement gains for traditionally disadvantaged student populations and other groups. School, district, and state accountability is the subject of debate on the local, state, and federal levels and serves as a call for schools to adopt reform efforts.

A great deal of scrutiny regarding high-stakes accountability systems and its counterproductive effects exists. Fullan (2011) asserted that such accountability produces a negative attitude in teachers and schools and creates a destructive effect on a school culture, "assuming that educators will respond to these prods by putting the effort to make the necessary change" (p. 8). In contrast to the negative impact that high-stakes accountability has on reform, Fullan identified four drivers of change necessary to judge the effectiveness and sustainability of reform efforts: fostering intrinsic motivation of

teachers, engaging educators and students in continuous instructional improvement efforts, inspiring collective team work, and affecting 100 percent of students in the effort.

In her identification of NCLB accountability as an impediment to high school reform efforts, Darling-Hammond (2006) identified four elements present in high performing urban schools: program personalization, well-qualified teachers, use of a common set of core academic standards, and targeted supports for struggling students. She noted that, "complicated rules that accompany NCLB have unintentionally made it more difficult for many heroic high schools in low-income neighborhoods to do their work well and keep the neediest students in school and moving to productive futures" (Darling-Hammond, 2006, p. 646). She suggested five problems with current NCLB legislation (Darling-Hammond, 2006). First, stringent NCLB definitions of highly qualified teachers make it difficult to recruit and retain quality teachers. Second, inconsistencies between state testing and accountability systems make calibrating efforts across states and districts difficult. Third, current reading and mathematics benchmark requirements in NCLB create disincentives for a school's efforts to use alternate forms of performance assessment to gauge student learning as many states use norm-referenced assessments. A fourth issue related to NCLB is the "diversity penalty" that schools containing large numbers of minority, special education, or limited English speaking students are required to meet all subgroup targets to be labeled as successful. A fifth shortcoming of NCLB pertains to practices by schools to withdraw poor performing students prior to testing to avoid accountability. These practices encourage pushing out students in efforts to raise test scores. Darling-Hammond (2006) and Ravitch (2011)

called for repealing NCLB and identifying ways in which to support instructional innovation for America's neediest students.

Structures that support reform. To make reform successful amidst the turmoil and politically-charged NCLB landscape, interdependence is necessary between school structures and instructional practices to take reform efforts to scale (McLaughlin, 1990; Oxley & Kassissieh, 2008; Quint, 2006). Reform efforts that target students most in need experience relatively few hurdles to implementation. Reforms focused on whole schools identified under NCLB regulations tend to encounter many more difficulties. Reorganization status may debase not only the school and staff, but the reforms that are applied to it. Oxley and Luers (2010) studied the progress of the federal Smaller Learning Communities (SLC) program and noted that districts that proactively launched reform initiatives across all high schools, regardless of their accountability status, conveyed the message that reform initiatives are a set of best practices for all schools as opposed to interventions suitable only for low performers thereby generating more favorable prospects for implementation.

Stringfield et al., (2008) noted that high reliability schools (HRS), like complex social organizations are "required to work under the very unusual demand of functioning correctly 'the first time every time'" (p. 412). HRS established finite goals, standardized operating procedures (SOP), and utilized data and data analysis to create a context in which failure is unacceptable. Fullan et al. (2004) identified the importance of finding appropriate structures that give districts a common direction and collective purpose, focusing on improving teaching and learning for both adults and students, and providing role clarity. McLaughlin (1990) called for revisions in existing federally-funded reform

programs in order to, "provide resources and support professional growth" (p. 13). In these instances, researchers noted that structural considerations are necessary at the school, district, and legislative levels to create and sustain innovation. The challenge, however, is to move quickly from the work of creating structures to begin close examination and refinement of school and classroom practices due to the fact that changes in instructional practices are widely regarded as the best way in which to raise student achievement levels (Bryk, 2010; Datnow, 2005; Fullan, 2000; Fullan & Pomfret, 1977; Quint, 2006).

Internal and external supports. To ensure that schools meet accountability benchmarks and have structures in place to support and sustain high school reform, district leaders must utilize a wide variety of internal and external supports. Internal supports include actions taken by an individual school or district to promote the reform effort and implementation. Fullan et al. (2004) called for school districts to build coalitions of leaders who have the ability to engineer reform and increase engagement with stakeholders through effective two-way communication. Datnow (2005) cited the importance of institutionalizing reform and the multiple factors that lead to the stability of the effort, noting that, "forces at the state, district, design team, school, and classroom level all interacted to shape the longevity of the reform" (p. 145). Full and sustained reform requires district stewardship that promotes a strong vision of instruction, focuses on a strong instructional core, shifts resources to support change, and ensures swift implementation (Oxley & Luers, 2010). The internal capacity of a school district to build, support, and sustain reform efforts is key to its ultimate success (Bryk, 2010; Datnow, 2005; Fullan, 2000; Fullan et al., 2004; Holdzkom, 2002; McLaughlin, 1990).

In their seminal work on systemic nature of school reform, Smith and O'Day (1991) identified three waves of educational reform. The first wave focused on "top down" reforms in the mid 1980's involving changes in educational inputs and a focus students demonstrating learning basic skills initiated by district officials. In the late 1980's, a second wave of reform surfaced that focused on, "decentralization, professionalism, and bottom-up change key concepts, as reformers focus on the change process and on active involvement of those closest to the reform" (p. 234). These reforms centered on the school as the individual unit of change. Smith and O'Day (1981) observed limitations in both of these approaches and proposed a third wave of reform that combined both district- and school-initiated reform activities. The researchers called for a "coherent systemic strategy" that takes place in a majority of schools within a district.

In a narrative synthesis of 81 peer reviewed research articles on district reform, Rorrer et al. (2008) identified four essential roles of school districts to support school reform. First, the authors found that districts provide instructional leadership that generates the will and capacity of reform for all schools in a district. As a second function, school districts reoriented the organization, refined organizational structures and processes, and made changes to district culture. Establishing policy coherence is a third function of school districts that involves managing federal, state, and local policies in addition to aligning district resources. A fourth role, maintaining an equity focus, involves a district's work to own and identify inequities within a district and establishing practices that promote accessibility and transparency for all schools within a district.

Adopting an existing reform model is a common approach that districts and individual schools take to seek improvements in student achievement. External partners

help districts build internal capacity for reform; provide strategies, resources, feedback, and professional development; and are subject to public review and scrutiny (American Institutes for Research 1999, 2006; Fullan et al., 2004; Fullan & Pomphret, 1977). In their review of scale-up efforts, Bodilly, Glennan, Kerr, and Galegher (2004) noted that technical assistance providers that provide education improvement services are relatively young and, "provide only limited evidence of their value and have only limited capacity to deliver high quality services" (p. 2). While the supports received through partnerships with technical assistance providers was an effective and popular means to accomplish whole school or district reform, the ability of a reform effort to adapt to the unique context of the schools served was equally important.

A delicate balance between reform fidelity and sensitivity to the individual context of a school makes full redesign effort implementation a challenge. Datnow (2005) noted that successful reform designs institutionalize reform involving, "a multilevel process of embedding an innovation in the structure and norms of the organization" (p. 123). The ability of a school or district to operationalize the key elements of a reform effort while adapting to the unique context in which a school operates was critical to the success of a reform (Aiken, 1942; Datnow, 2005; Datnow & Stringfield, 2000; Fullan & Pomphret, 1977; Holdzkom, 2002; Stringfield et al., 2008). In addition to a school district's adaptability and flexibility, their ability to manage the supports and activities from a variety of stakeholders was critical for scale up success. Bodilly et al. (2004) noted that, "if scale-up is to succeed, the actors involved—including developers, district officials, school leaders, and teachers—must jointly address a set of

known, interconnected tasks, especially aligning policies and infrastructure in coherent ways to sustain practice" (p. 648).

Problems Taking Reform to Scale

Sustaining high school redesign efforts in a politically-charged and turbulent context is difficult for districts wishing to improve schools and sustain public confidence. Current NCLB and School Improvement Grant (SIG) guidelines require quick improvements in order to avoid sanctions that require removing principals and teachers, relinquishing control of the school to an external management organization, adopting a performance-based transformation model, or closing persistently low-achieving schools. The disconnect between the time required to take a reform effort to scale and the time mandated to improve is a key factor in the failure of reform efforts going to scale. Fullan (2000) noted that high school reforms take five to six years to take hold, while district-wide efforts take six to eight years. In his review of Comprehensive School Reform (CSR) projects, Holdzkom (2002) observed that gains following a school reform effort became evident after three years of implementation. High stakes accountability systems that require quick gains hinder schools implementing reforms by forcing schools to put reform aside for test preparation, placing a premium on instructional strategies versus deep reform models, eliminating programs that may be of great benefit to students but not measured on accountability tests, and creating rules and policies that stymie innovation (Darling-Hammond, 2006; Datnow, 2005; Holdzkom, 2002).

In their definition and study of "third age" school improvement efforts in Kentucky, California, Chicago, New Zealand, and Australia, Hopkins and Reynolds (2010) noted that modern performance-based reforms have done little to improve the core

technology of schools. Hopkins and Reynolds noted that first and second age reform efforts were "free-floating" initiatives that focused on organizational change, targeted accountability, and promoted value-added measures for judging schools. The authors cautioned that in the absence of valid reform designs, school leaders risk selecting previously failed reform models that lack systematic approaches to change and promote turn-key practices that prevent teachers from generating innovation internally. Well-intentioned school districts implementing reforms that do not place improving instruction at the center of attention risk spending time, resources, and political capital on futile efforts that do little to advance student achievement.

Although providing schools and districts with the time necessary to implement reform efforts is necessary, developing the capacity in schools to execute new innovations was of equal importance. Bryk (2009) identified the lack of infrastructure to guide transformation at the school and district level. He called for the need to, "engineer both how we carry out education R & D and the institutional environments in which this work occurs if we want to achieve more productive ends" (p. 597). Bryk (2009) recommended the use of a Design-Education Engineering-Development (D-EE-D) framework to carry out quick and effective changes in day-to-day instructional practices in classrooms and schools. As part of the D-EE-D framework, Bryk advocated a rapid prototyping process consisting of developing instructional innovations, trying them in schools, and refining practices based on teacher feedback and academic results. According to the author, D-EE-D, focuses on day-to-day instructional practices and merges the scientific discipline of action research and systemic approaches.

Instructional Program Coherence

Elements of school reform driven by purpose, implementation structures, and effective internal and external supports have been necessary and at times successful, but not reliably sufficient to move significant numbers of students to proficient performance in urban school districts (Earl, Torrance, & Sutherland, 2006; Payne, 2008; Stringfield & Datnow, 1998). Genuine and sustainable reform may require coherence of the elements of reform through an overarching strategy (Childress, Elmore, Grossman, & Akinola, 2004). Newmann, Smith, Allensworth, and Bryk (2001) asserted that initiatives of curriculum alignment for grade-to-grade transition, school organization for unity of purpose, whole-school design for reform connectedness, and program coordination with district and state policies have produced pockets of improvements, but have typically lost momentum and sustainability. Districts fall into a "fragmented circuit of school improvement activity" (Newman et al., p. 298). Instead of a variety of programmatic changes, schools need instructional program coherence to coordinate structures, staff working conditions, and resources uniformly aimed at improving student achievement (Honig & Hatch, 2004; Kedro, 2004; Newmann et al., 2001; Oxley, 2008).

Newmann et al. (2001) defined instructional program coherence as "a set of interrelated programs for students and staff that are guided by a common framework for curriculum, instruction, assessment, and learning climate and are pursued over a sustained period" (p. 299). Through teacher surveys, student test scores, and field studies within 222 Chicago elementary schools, Newmann et al. found schools that improved instructional coherence also improved student achievement. From their study, three conditions surfaced as evidence for improved instructional coherence: a common

instructional framework for guiding teaching and learning, staff working conditions to support implementation of the common framework, and coordinated resources to support the framework.

Project Proficiency

The Jefferson County Public Schools (JCPS), a large urban district of approximately 100,000 students in Louisville, Kentucky, has created a system, Project Proficiency (PP), that is designed to meet the Newmann et al. (2001) litmus test for robust, instructional program coherence. Results from the 2010-2011 school year indicated that all 21 JCPS high schools gained in reading and math proficiency, with the 11 PLA and near-PLA schools averaged a 14% gain in reading and a 17% gain in math, tripling state gains (Kentucky Department of Education, 2011).

To connect practitioners and coordinate reform efforts amid the landscape of NCLB sanctions and challenges of advancing disadvantaged high school students inherent with large urban districts, PP boldly established an overarching strategy of "guaranteed competency," or the goal of ensuring learning of key reading and math standards by each student. Through the "strategic function" (Childress, Elmore, & Grossman, 2006, p. 59) of guaranteed competency, JCPS created district-wide instructional program coherence evidenced by a common instructional framework, complementary staff working conditions, and supportive resources.

Common instructional framework. Through a narrow curriculum, balanced assessment system, and purpose-driven instructional principles, PP enabled teachers to guarantee student competency of three key standards each grading period and leverage a coherent common instructional framework. First, each six weeks, the district established

three priority standards with corresponding curriculum maps for core high school English and math courses, providing clear learning targets and expectations for staff and students. Unifying schools and the district around a reduced and nonnegotiable set of content standards provided a common direction for students, school staff, and district administrators. These goals resembled the set of goals that characterized highly reliable organizations (HRO) (Datnow & Stringfield, 2000; Stringfield et al., 2008).

Second, guaranteeing competency produced a "balanced assessment system" (Stiggins, 2008, p. 3) to track student progress. PP included a district diagnostic assessment to determine early levels of student understanding of each standard, a summative proficiency assessment for an end-of-grading-period measure, and frequent teacher-designed formative assessments to evaluate student improvement toward competency. Basing student grades on demonstrations of competency unified teachers around standards-based instruction and assessment. Teachers reinforced their standards-based approach with opportunities for students to reflect on their own progress, cited by Stiggins (2008) to positively impact student achievement. To ensure learning, teachers were required to guide each student to demonstrate competency for each of the three key standards by the time the proficiency assessment was administered, and those scoring below 80% on the proficiency assessment were guided to recover or correct their work until they met the threshold target. Through a balance of diagnostic, formative, and summative assessments designed to guarantee student competency of clear standards, PP converted high school assignments, tests, and grades into a coherent system to ensure learning.

Third, similar to the Coalition of Essential Schools design to improve teaching

and learning through guiding principles rather than a packaged program (Coalition of Essential Schools, 2011), PP coalesced instructors around the precepts of shared accountability for high-level, standards-based teaching, and ownership of student results. To guarantee competency, teachers needed to create tasks through which students could demonstrate understanding, develop lessons with focused learning targets aligned with those tasks, and ensure each student demonstrated understanding of each key standard. With instruction tied to required outcomes, teachers regularly adjusted how they delivered instruction, assessed, and intervened for struggling and reluctant learners. Guaranteeing competency transformed teachers from "directors into diagnosticians" (Kedro, 2004, p. 32), shifted their mission from ownership of teaching to ownership of learning, and merged curriculum, assessment, and instructional systems into a seamless, coherent, and common instructional framework.

Working conditions. Complementing a common instructional framework, PP fostered working conditions characterized by collective teacher efficacy (DuFour, DuFour, Eaker, & Karhanek, 2004; Newmann et al., 2001). Establishing the goal of guaranteed student competency generated levels of collaborative practice, decision-making, and professional development not previously experienced by teachers. Due to the goal of moving each student to levels of competency by the end of each six weeks, teachers of common courses met weekly and sometimes daily to diagnose learning gaps and exchange updates on the numbers of students meeting competency. Drawing ideas from one another, instructors collectively designed new lessons, tasks, and interventions to address student deficiencies. District resource teachers provided recommendations for adjustments and ideas from other teacher teams. School-based administrators promptly

responded to teacher requests for time, resources, and support. Collaborative reflection, collective action, and collegial "expertise development" (Bryk, 2009, p. 599) produced a coherent learning climate for practitioners through their "agency that produced the texts, rules, and guidelines of their school change process" (Stringfield & Datnow, 2002, p. 282).

Coordinated resources. JCPS completed its coherent instructional program design with unprecedented support resources of curricular materials, data management, and principal leadership. Childress, Elmore, and Grossman (2006) asserted, "Only the district office can create such a plan, identify and spread best practices, develop leadership capabilities at all levels, build information systems to monitor student improvement, and hold people accountable for results" (p. 55). With input from local school practitioners, district curriculum specialists identified the key standards for each core course from state mandated content, developed curriculum pacing guides for each grading period, and designed corresponding diagnostic and proficiency assessments.

To provide effective and timely student performance information and positively impact interventions at the classroom, school, and district levels (Stiggins & DuFour, 2009), JCPS designed a web-based data entry system for diagnostic, formative, and summative assessment results that provided teachers with details for tracking student demonstration of competency, diagnosing possible content misunderstandings, and converting standards-based evaluation of student competency into grades (Jefferson County Public Schools, 2011a). Leadership made the ultimate difference for effective supervision and support. Principals provided common planning and facilitated teacher learning team protocols to foster collective efficacy; district and state improvement

funding afforded additional materials, staff, and stipends; and district leaders organized principals into accountability teams for comparing school data trends, exchanging leadership innovations, and assessing district instructional needs.

Promoting coherence through a district-wide strategy of guaranteeing student competency of key standards, JCPS implemented PP to move dramatic numbers of high school students to proficient performance in one school year. However, institutionalizing this reform across its urban high schools confronted JCPS with a formidable challenge. Fullan (2001) asserted, "twenty-five percent of the solution is having good directional ideas; 75% is figuring out how to get there in one local context after another" (p. 268). Having met the criteria for instructional program coherence, JCPS needed to move its PP reform to scale at the district level.

Rethinking Scale

Over the past decade, educators across America have focused on turning around low performing schools (Forte, 2010; Linn, 2009; Ravitch, 2010). Reform efforts in schools and districts have included replacing principals and teachers, transferring authority of schools to outside management organizations, and even more drastic, closing schools (School Improvement Fund, 2010). Finding ways to turnaround low performing schools and districts has dominated recent discussions in education circles. Little evidence exists to support successful reform in high schools as most of the evidence reporting successful turnaround and reform exists at the primary school level (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010; Stringfield et al., 2008; Slavin et al., 1996). This lack of success especially occurs within urban high schools and districts. Since the inception of the persistently low achieving label in Race to the Top legislation,

urban high schools have dominated states' lists of lowest performing schools. In Kentucky alone, 41 schools have been identified as persistently low achieving since 2009. Of those 41 schools, 33 of these lowest performing schools are high schools and nearly 40% of those schools, are urban high schools.

Although reform at the high school level has seen limited success on the individual school level, taking urban high school reform to scale at the district, state, or even national level has proven unsuccessful (Payne, 2008; Stringfield & Datnow, 1998). From the Coleman Report (Coleman et al., 1966) to NCLB legislation to Race to the Top legislation, education policymakers seek ways to take high school reform to scale. Despite this desire to reform high school education in America, sustainability is "extremely rare" (Stringfield et al., 2008). Stringfield and Datnow (1998) define scale as "the deliberate expansion to many settings of externally developed school restructuring design that previously has been used successfully in one or a small number of settings" (p. 271). In order for education reform to impact schools across America, educators must find ways to bring urban high school reform to scale. Urban districts must implement reform efforts that do not depend on individual school improvement efforts, but instead take these efforts to scale across a large number of high schools.

Coburn (2003) provided a framework for successful reform in education. Coburn not only provided the framework for reform efforts in education, but more importantly, the necessary factors to move this reform to scale over a large number of schools. She detailed four major dimensions necessary to accomplish this feat. According to Coburn, the four interrelated dimensions critical to moving educational reform to scale included depth, spread, shift of ownership, and sustainability. If any one of these four dimensions

are missing, reform efforts cannot be taken to scale and eventually will fail. As previously mentioned, JCPS implemented Project Proficiency (PP) to bring reform to scale with reading and mathematics instruction and to move dramatic numbers of students to proficiency in a short time. PP had several components that addressed the framework provided by Coburn to bring this reform to scale.

Depth. According to Coburn (2003), most reform efforts in secondary education do not substantially change the instruction that occurs within every classroom. Any reform efforts must, however, consist of deep change that drastically alters classroom instruction.

> By "deep change," I mean change that goes beyond surface structures or procedures (such as change in materials, classroom organization, or the addition of specific activities) to alter teachers' beliefs, norms of social interactions, and pedagogical principals as enacted in the curriculum (p. 4).

Coburn directly referred to the depth of change necessary to bring reform to scale. Teachers must rethink or reconstruct their views and beliefs about instruction in order to make this consequential change in classroom practice.

Unlike previous JCPS reform efforts, PP attempted to foster deep pedagogical change in teacher grading practices to move reform to scale. Historically, JCPS teachers graded students on a traditional point system. Teachers gave students grades based upon point accumulation through in-class assignments, homework, quizzes, tests and projects. With the implementation of PP, teachers pursued the goal to move all students to proficiency in three key standards during a six-week grading period. With this movement towards standards-based grading, teachers had to reconstruct their views and beliefs

about grading practices. Teachers graded students exclusively on achievement of key standards instead of other factors such as growth, effort, behavior, homework completion, and attendance.

Without the implementation of standards-based grading, the goal of ensuring competency in the three key standards could not be met. The shift in grading practices caused teachers to change their pedagogical beliefs. Ensuring competency in standards required teachers to go beyond "surface structures or procedures" (Coburn, 2003, p. 4) by collaborating with colleagues on a regular basis to provide innovative approaches that bring all students to proficiency. Teachers could measure whether students had met proficiency standards only through the use of a standards-based grading approach. The move to standards-based grading caused a unique depth of change in secondary mathematics classrooms never experienced in JCPS high schools.

Spread. In order to bring reform to scale, leaders must both spread the reform across classrooms and schools and spread the underlying beliefs to as many schools and classrooms as possible. According to Coburn, "Rather than thinking of spread solely in terms of expanding outward to more and more schools and classrooms, this emphasis on the normative highlights the potential to spread reform-related norms and pedagogical principles within a classroom, school and district" (Coburn, 2003, p. 7). The implementation of a reform must occur across many schools and classrooms, but the "underlying beliefs, norms, and principles" must also be present for the reform to be brought to scale.

Eleven of the 21 JCPS comprehensive high schools were already under state sanctions for persistently low-achieving status or in serious danger of being identified as

one. These schools also had some of the lowest proficiency rates in mathematics. In order to prevent future state sanctions, reform efforts in mathematics had to be brought to scale in a large number of schools and classrooms. Eleven high schools implemented PP in every Algebra 1, Geometry, and Algebra 2 classroom. But, implementation of reform across a large number of schools and classrooms was not sufficient; the underlying "beliefs, norms, and principles" concerning instruction had to move to all classrooms within each of the schools. No longer could large variation in mathematics instruction occur from classroom to classroom and school to school. The principles behind PP reduced this variance by ensuring competency in each of the three key standards. Teachers had to align their instruction for proficiency in standards by collaborating daily with colleagues in order to be successful (DuFour et al., 2004). In previous years, proficiency rates differed between classrooms and schools. Now with PP, all teachers had to work together in order to ensure learning of key standards for each grading period.

Shift of ownership. Bringing reform to scale required a shift in ownership of the reform. Too often ownership of the reform lay outside of the school and classroom. Coburn (2003) noted, "One of the key components of taking reform to scale, then, is creating conditions to shift authority and knowledge of the reform actors to teachers, schools, and districts" (p.7). In short, teachers and schools must truly buy-in to the reform and customize it for each classroom. By allowing teachers and schools to create and operationalize reform to meet schools needs, authority shifts from the district level to the classroom level (Datnow, 2005).

Within the PP framework, teachers had to ensure the proficiency of every student in three key standards for each grading period. The framework created conditions where

individual teachers and schools customized interventions to ensure the learning of each standard. Unlike previous instruction where content was covered, grades were assigned, and the scope and sequence of the course continued, teachers were charged with ensuring learning for all students. Groups of teachers collaborated to bring every student to competency in each standard and owned the reform effort and the requirement to ensure learning of key standards. As a result, the reform shifted from district and even school ownership to teacher and classroom ownership.

Sustainability. Finally, sustainability is a key factor in bringing reform to scale. According to Coburn, "the concept of scale primarily has meaning over time. The distribution and adoption of an innovation are only significant if its use can be sustained in its original and even subsequent schools" (Coburn, 2003, p. 6). Sustainability often becomes the most challenging dimension in bringing reform to scale. Too often, districts and schools change priorities and demands or face change in staff and leadership. Therefore, reforms are often short term and result in a lack of scalability.

PP completed its first year of implementation during the 2010-11 school year. If successful, the challenge for the district and schools involved becomes sustainability. Future success of PP requires continued district and school support of the reform along with the continued spread throughout new schools and subject areas. The purpose of this capstone was to examine if PP impacted student achievement and contained conditions for taking the reform to scale. To achieve this purpose, the researchers studied PP through investigations of whether a standards-based grading approach was associated with higher achievement, gains in student outcomes remained after controlling for

classroom and school compositional effects, and PP's impacted achievement of the district's students who were most at-risk of dropping out.

THE ASSOCIATION BETWEEN GRADES AND TEST SCORES IN STANDARDS-BASED GRADING

A movement has emerged in secondary education to grade students solely on achievement in key academic standards within a curriculum (Guskey, 2009; Marzano, 2010). Guskey (2009) describes standards as "what we expect students to know and be able to do as a result of their experiences in school" (p. 1). Standards became a central focus in education with the implementation of No Child Left Behind (2001). Race to the Top (2010) competition awarded states points for the implementation of common core standards. Despite this focus on standards-based reform, little evidence exists that secondary classroom teacher grading practices have substantially changed to measure student achievement in terms of these common standards (Guskey, 2007). Even with the new focus on standards-based grading, most secondary teachers use a "hodgepodge" of factors other than achievement in key standards to determine students' grades (Brookhart, 1993; Cross & Frary, 1999; Guskey, 2007; McMillan, 2001; Stiggins, Frisbie, & Griswold, 1989). For example, teachers rely on assessment of non-standards processes such as effort, behavior, class participation, homework completion, ability level, and growth (Brookhart, 1993; Cross & Frary, 1999; Guskey, 2007).

With the current atmosphere of high stakes accountability, administrators and teachers are held accountable for ensuring learning of these standards within their classrooms and schools. In order to ensure the learning of these standards, teachers must

align student grades with mastery of key academic standards within the core content. Since the shift to standards-based grading calls for deep and systematic change to long lasting educational traditions, clear demonstrations of its benefits are needed.

The purpose of this study was to determine whether the implementation of a standards-based grading approach was associated with higher achievement on standardized tests as compared to traditional grading practices. Specifically, this study was designed to answer the following research questions:

1. Does a stronger association exist between standards-based grading and student achievement than with traditional grading practices and student achievement?
2. Does a stronger association exist between standards-based grading and minority student achievement than with traditional grading practices and minority student achievement?
3. Does a stronger association exist between standards-based grading and at-risk student achievement than with traditional grading practices and at-risk student achievement?

Background

Over the past decade, United States domestic policy has included a clear focus on reform in public education. This reform attempts to change many long standing traditions within secondary education. Despite these reform efforts, teachers and administrators have not significantly changed the practice of grading. Cizek, Fitzgerald, and Rachor (1996) observed, "It seems that classroom assessment practices may be a weak link in the drive toward improving American education" (p. 162).

Federal and state governments now hold schools accountable for student learning of specific standards. Students are expected to demonstrate proficiency of these standards on standardized state accountability assessments. Therefore, the most critical component of educational reform may be the implementation of sound grading practices

that directly measure student attainment of required standards. Despite this necessity for change in assessment, many teachers do not measure students solely on their achievement level according to standards, but instead measure subjectively on a variety of factors that cloud the essential fact of whether students have met standards. In order to ensure a high quality education for all students, an education in which no child is actually left behind, reform in grading must become pervasive throughout secondary education in America. As Guskey (2009) stated,

> If grades are to represent information about the adequacy of students' performance with respect to clear learning standards, then the evidence used in determining grades must denote what students have learned and are able to do. To allow other factors to influence students' grades or to maintain policies that detract from that purpose misrepresents students' learning attainment (p. 22).

Hodgepodge of Grades

To make systemic change within secondary education, measurement researchers state that grades need to be based solely on levels of achievement within a class (Allen, 2005; Cross & Frary, 1999; Guskey, 2009). The vast majority of prior research on grading in secondary education indicates that most teachers fail to grade on achievement. Brookhart (1991) initially described the grading process in secondary schools as a "hodgepodge of attitude, effort, and achievement" (p. 36). Most of these research studies involved surveying teachers on the various factors that they take into account when giving a student grade in their class. Brookhart (1993) also found that the 84 teachers surveyed used the image of grades as classroom currency to encourage student effort, participation, and appropriate behavior within the classroom. Teachers within

Brookhart's research clearly felt that both academic achievement and effort are relevant when grading a student.

Cross and Frary (1999) further explored Brookhart's findings of the hodgepodge of factors used in the grading of secondary students. On the basis of their survey of 307 teachers, the researchers confirmed teachers' use of many non-achievement based factors in grading students when they concluded:

> Because of the importance placed on academic grades at the secondary level, either for educational or occupational decisions, grades should communicate as objectively as possible the levels of educational attainment in the subject. To encourage anything less, in our opinion, is to distort the meaning of grades as measures of academic achievement, at a time when the need for clarity of meaning is greatest (Cross & Frary, 1999, p.56).

McMillan and Nash (2000) further investigated the influences on teacher decision making with respect to grading and the justification that teachers give when assigning grades. The authors surveyed 700 teachers and interviewed a sample of these teachers; then they coded their responses concerning various classroom factors involved in grading. Although achievement, as defined by student understanding, was one of the main categories they identified on the basis of their research, several other categories emerged. These categories included the teacher's philosophy of teaching and learning, the teacher's desire to "pull for students," the teacher's accommodating individual differences among students, and finally student engagement and motivation. Throughout their research, McMillan and Nash found that teachers used grades as the main tool to encourage and monitor student engagement. Although teachers verbalized the need to

measure student achievement through grading, "most teachers used a variety of assessments….including homework, quizzes, tests, performance assessments and participation" (McMillan & Nash, 2001, p. 26).

To better understand and explore the various factors used in grading, McMillan (2001) surveyed 1483 teachers and identified four distinct "hodgepodge of factors" most often seen in secondary grading practices. These factors included academic achievement, external benchmarks, academic enablers, and extra credit. Although McMillan found significant evidence of the use of academic achievement in grading, the other three factors weighed heavily in teacher's assessment. McMillan also found teachers assessed higher ability students in a motivating and engaging environment by measuring higher cognitive skills, while the same teachers gave lower ability students more rote learning assessments, more extra credit and less emphasis on academic achievement. This finding supports the belief that grading practices in secondary schools increases the achievement gap that exists between many subgroups of students. Higher ability students are graded based upon achievement while many at-risk students indentified in subgroups are graded on a wider range of factors.

Validity of Grades

The results of the survey research of secondary teachers' grading practices show that teachers use a "hodgepodge" of factors to grade students. Student achievement emerged as only one of the factors used by teachers to assess student work; therefore, grades are not necessarily a valid measure of students' level of achievement in secondary education. Educators make critical decisions about the future of students by using grades that are supposed to measure student achievement. These decisions include entry into

elite clubs and organizations, access to scholarships, and admissions into college. If grades measure several factors, including a student's ability to navigate the social processes of school and not just student achievement, the validity of grades becomes a major concern in American education. For grades to be a valid measure of student achievement, teachers must assess students on their achievement based on required curriculum standards.

As a result of the variety of factors used by teachers to grade students, Marzano (2000) contends that in terms of measuring student achievement "grades are so imprecise that they are almost meaningless" (p. 1). Allen (2005) summarizes the critical nature of ensuring validity in the grading process in measuring academic achievement.

> Also, since many of these factors such as effort, motivation, and student attitude are subjective measures made by a teacher, their inclusion in a grade related to academic achievement increases the chance for the grade to be biased or unreliable, and thus invalid. The purpose of an academic report is to communicate the level of academic achievement that a student has developed over a course of study. Therefore, the sole purpose of a grade on an academic report, if it is to be a valid source of information, is to communicate the academic achievement of a student (p. 220).

Guskey (2007) explored the perceived validity of teacher grades by surveying 314 educators in three different states. He asked educators to rank from 1 to 15 sources of evidence of student learning that "you trust to best show what students know and can do" (p. 21). The sources of evidence included standardized tests, various assessments, teacher observations, quizzes, homework completion, portfolio, students' grades, class

involvement, and behavior and attitude. Statistical analyses of the data from the study indicated that the participants gave a relatively low ranking to grades being an accurate indicator of student learning. Guskey (2007) contends that the educators' low ranking of grades correlating with academic achievement results from both teachers' and administrators' recognition "that a variety of nonacademic factors, such as effort, attitude, participation, and class behavior, typically influence grades" (p. 22). These factors also support the discrepancy found between student grades and standardized test scores (Allen, 2005).

In another study on the validity of grades, Bowers (2009) explored the relationship between teacher assigned-grades and standardized assessments. He found that schools use standardized test scores to make data-driven decisions, in place of grades. Administrators have consistently sought to remediate and intervene for low performance on standardized tests, when student grades should also be used to inform these decisions. Since grades are not just a valid measure of a student's achievement level, schools can use this information to intervene and assist students to improve on the various factors that grades do measure.

> The hypothesis here is that rather than cast this hodgepodge nature of grades in pejorative light as data that is useful to schools because grades only moderately correlate to test scores, the theory presented here...points to the idea that grades appear to assess both academic knowledge…as well as a student's ability to perform well at the social tasks of the schooling process, such as behavior, participation, and attendance (Bowers, 2008, p. 622).

Bowers conceded that grades are not a valid measure of a student's academic achievement; therefore, schools must use the basis of grades to provide critical safety nets to support student success.

Relationship Between Grades and Test Scores

A substantial amount of empirical research does not exist on the relationship between grades and standardized test scores. Little evidence exists on the impact of standards-based grading on standardized test scores. As a result of schools' increased accountability for improving standardized test scores, several research studies have attempted to determine the relationship between grades and test scores. If recent increases in school accountability have led to changes in teacher grading practices, then an association should exist between grades and standardized test scores. If the use of standards-based grading methods has led to a decrease in the use of a hodgepodge of factors to assess student learning and grades in turn are more of a valid indicator of student achievement, then a strong correlation should exist between grades and test scores.

David Conley (2000) examined the relationship between grades teachers give their students and proficiency scores given to the same students by external raters. Conley found a lack of relationship between teachers' grading system and student proficiency. He specifically noted that students judged proficient through an analysis of their work by external raters were not necessarily the students with high grades. "The stepwise regression analysis examines teacher grading systems and student proficiency scores and found very little relationship between the grading system a teacher used and whether or not a student was proficient" (p. 18). Conley surmised that the low

correlation suggests that separate constructs besides standards based achievement are used in grading. Specifically, he noted that homework in math classes and in-class assignments in English classes make up a significant portion of a student's grade, although these assignments might not measure proficiency on mandated standards.

This relationship between test scores and grades has been topic of several research studies over the past decade. Lekholm and Cliffordson (2008) studied the grades of nearly 100,000 students from Sweden and their association with students' scores on national tests. Although results from their analysis indicated that the greatest variance in grades came from actual achievement levels in the subject area, other factors outside of achievement influenced the grades given to students. One of the most significant findings of their research revealed that schools with students from lower levels of socio-economic backgrounds assigned grades that were higher than their standardized test scores. Therefore, the at-risk students in these schools presented evidence of a lower correlation between grades and test scores.

Two empirical studies discovered only modest correlations between teacher assigned grades and standardized state assessments. Brennan, Kim, Wein-Gross, and Sipperstein (2001) as well as Happonstall (2010) not only explored the correlation between grades and standardized tests, but also differences in the association between grades and test scores for minority, low socio-economic, and non-minority students. Both studies found a lower correlation between grades and standardized test scores for minority students, English language learners and lower socio-economic students than their counterparts. The findings suggest not only that grades do not strongly correlate with achievement scores on standardized tests, but that minority students and low socio-

economic students are possibly given higher grades than their achievement levels warrant. These findings support a theory of grade inflation with these at-risk students as a result of teachers using a "hodgepodge" of factors such as effort, behavior and attendance to justify high grades. According to Brennan et al.(2001) and Happonstall (2010), the practice of grading at-risk students on factors other than achievement level supports the existence of a significant achievement gap between minority students and their white counterparts. Despite intense focus on the elimination of the achievement gap in American secondary schools, few education leaders have examined grading policies as a potentially significant component of the problem.

Based on two decades of research on the grading practices of teachers in secondary schools, researchers found that teachers are evaluating students on a hodgepodge of factors that do not validly assess a student's achievement level in a specific content area. Such grading practices have potentially increased the achievement gap in American education because of the inflation of grades for at-risk students. Research data is limited on the correlation between grades and achievement scores on standardized tests at the secondary level. Even less research is available on the impact of standards-based grading on the correlation between grades and achievement scores. The current study examined the association between standards-based grading and achievement as measured by standardized tests. The authors examined effects, for students in general and for minority and lower socio- economic students in particular.

Method

Participants

This study included participants from 11 high schools that implemented a district designed program named Project Proficiency (PP) for the 2010-11 school year. Educators implemented PP in order to improve proficiency scores on the Kentucky Core Content Test (KCCT) in mathematics. All 11 high schools are a part of Jefferson County Public Schools (JCPS) in Louisville, Kentucky. JCPS, the 26th largest school district in the nation, serves 100,474 students in 150 schools. The demographic composition of the district's student body is 51% white, 37% African American and 12% other students. Nearly 62% of the district's students qualify for the federal free/reduced lunch program. Table 1 provides demographic information for each cohort.

Table 1

Cohort Characteristics (N = 2419)

	Non-PP-Cohort (2009-10)		PP-Cohort (2010-11)	
Characteristic	n	%	n	%
Gender				
Male	585	50.3	610	48.6
Female	578	49.7	646	51.4
Race/ethnicity				
Caucasian/White	509	43.8	552	43.9
Minority (non-white)	654	56.2	704	56.1
Free/reduced lunch	822	70.7	920	73.2
Total	1163	100	1256	100

From the high school population, the researchers drew two separate cohort groups. One cohort consisted of 11th grade students from the 11 high schools during the 2011 school year. Each of these students completed an Algebra 2 course and received PP within the Algebra 2 course. The second cohort consisted of 11th grade students from the same 11 high schools from the preceding (2009-10) school year. Because PP had no yet

been developed, those students did not receive PP within their Algebra 2 course. Students in this study completed an Algebra 2 course during the 2010 or 2011 school years and had grade 11 KCCT results in mathematics (11Math) and science (11Science) during the same year. In Kentucky, students take both the 11Math and 11Science assessment during the 11th grade. Juniors in the 11 high schools experienced PP in math but not in science. Therefore, science scores were used to provide a statistical control when comparing the effects of PP between the cohorts of students in their Algebra 2 courses.

Algebra 2. For purposes of this study, Algebra 2 was defined as a course in the Kentucky Program of Studies that meets the Kentucky Algebra 2 graduation requirement. These courses included Algebra 2, Algebra 2 Honors and Algebra 2 Advanced. These courses also included students who qualified for special education services. Students who did not complete an Algebra 2 course or who did not have a KCCT combination in mathematics and science during the same academic year were excluded from this sample. Algebra 2 classrooms were identified in each of the study's schools through an evaluation of each master schedule. All students taking Algebra 2 in the two cohorts and 11 schools were included in this study.

Measures

Student data for each of the cohorts were obtained from the Jefferson County Public Schools Data Warehouse. Kentucky administers the math and science test to students in grade 11, and for security reasons, distributes multiple versions of the test. The KCCT Test Administration Guide identifies average Chronbach's Alpha measures of .89 and .84 for the six versions of the mathematics and science tests. Item and

description indices were identified by the Kentucky Department of Education for each test version and converted to mean scale scores (MSS) from 0-80. Kentucky mean scale scores correlate to four performance level descriptors: novice, apprentice, proficient, and distinguished. Table 2 below describes scale scores and performance level descriptions for both 11th grade mathematics and science tests.

Table 2

Grade 11 Kentucky Core Content Test (KCCT) Mean Scale Score Range and Performance Descriptors

	Scale Score Range			
Content	Novice	Apprentice	Proficient	Distinguished
Mathematics	0-19	20-39	40-63	64-80
Science	0-19	20-39	40-62	63-80

For the purpose of this study, letter grades were converted to numerical scores. Table 3 shows the numerical scores for each of the letter grades.

Table 3

Letter Grade Value and Range

Letter Grade	Value	Range
A (Exceeds Standards)	4.0	93-100
B (Meets Standards)	3.0	86-92
C (Marginally Meets Standards)	2.0	79-85
D (Below Standards)	1.0	70-78
F/U (Unsatisfactory Performance)	0.0	0-69

Note. Descriptions and grade ranges come from Jefferson County Public Schools "Student Progression, Promotion and Grading"

Design and Procedures

A quasi-experimental nonequivalent control group design was implemented to analyze the association between standards-based grades and 11Math during the 2011

school year. Comparison of the association between grades and KCCT scores relied on two control groups. The first control group consisted of students who took Algebra 2 and received an 11Math score in 2010, but did not experience the standards-based grading effects within PP. The first control group provided a measure of association between two years in mathematics, and the second control group provided a comparison of association between two groups of the same students. The second level of control involved the same students as the treatment group. These students received standards-based grading as a part of PP in mathematics but not in science. These students also took 11Science, and the study analyzed the association between their grades in science and their 11Science scores. This association provided a direct comparison to the mathematics association. The second control allowed the researchers to remove limitations caused by demographic differences within the two cohorts.

Variables

Both 11th Grade Mathematics Mean Scale Score (11Math) and 11th Grade Science Mean Score (11Science) were used as dependent variables. 11Math was used as the primary dependent variable and 11Science was used as an additional dependent variable to compare the effects of PP within the same treatment group.

In this study, students' mathematics test scores were treated as the dependent variable. Specifically, the particular effect analyzed within this study was the association between standards-based grades within PP and 11Math scores. Standards-based grading in PP was only used with the 2011 cohort for mathematics. The 2010 cohort in mathematics and the 2011 cohort in science were evaluated and graded based on a traditional grading approach. The Jefferson County Public Schools Student Progression,

Promotion and Grading (SPP&G) (Jefferson County Public Schools, 2011b) provides the explanation for the traditional grading policy. The SPP&G defines district policy concerning the components of an academic grade. The policy states, "Academic grades must include a minimum of three of the following: portfolios, projects, discussion/problem solving, group work, classroom assignments, homework/journals/logs, quizzes, tests, participation, and teacher observation" (p. 8). Finally, district policy mandates that "one component may not count for more than 40 percent of the total academic grade" (p.8). Table 3 provides an explanation of academic grades.

The independent variable used for standards-based grading was the implementation of PP. Instead of using the traditional grading method, PP assessed students based on standards-based grading approach. As a part of the standards-based grading process in 2011, teachers required their Algebra 2 students to become proficient in three key standards for each six week grading period. Teachers graded students within the 2011 cohort in mathematics solely on their proficiency level for each of the key standards within the grading period. Students took both a diagnostic assessment in the middle of the grading period and a proficiency assessment at the end of the grading period. Each of these assessments accounted for 40 percent of the students' grade. Student reflection on their proficiency within each standard accounted for the final 20 percent (Jefferson County Public Schools, 2011a). "Web-based technology provides teachers with a detailed system for tracking student demonstration of competency, diagnosing possible content misunderstandings, and converting standards-based evaluation of student competency into grades" (Jefferson County Public Schools, 2011a,

p. 3). Finally, PP required that students who do not reach proficiency remediate after the grading period in order to meet the key standards and retake proficiency assessments. As a result, grades for mathematics in the 2011 cohort were based solely on a students' proficiency level in the three key standards taught in each of six, six-week periods.

Evaluation of Grades

First, the researchers used basic descriptive statistics to determine the percentage of students in each cohort that scored proficient or above and received an A or a B in the corresponding content course. On the KCCT assessment, proficient or above is the level necessary for students to score in order for schools to avoid state and federal sanctions.

Second, the researchers evaluated students who received above average grades within the content course of each of the three cohorts. Students who received an A or a B in the specific content course were considered above average in standard attainment. An analysis of variance (ANOVA) determined whether students who experienced standards-based grading and scored above average in their class scored higher on the corresponding KCCT assessment than students who experienced traditional grading. The one-way ANOVA compares the means of two or more groups of participants that vary on a single independent variable.

Association Statistics

The analysis determined the correlation coefficient (r) for each of the three groups. "The correlation coefficient is an index that describes the extent to which two sets of data are related; it is a measure of the relationship between two variables" (Hinkle, Wiersma, & Jurs, 2003, p. 98). By analyzing the correlation coefficient for each of the cohorts, the researchers determined whether the treatment group that received standards-

based grading had a higher correlation to KCCT math scores than the control groups who did not receive standards-based grading. In order to answer each of the research questions, coefficients of determination also were calculated in each cohort for both minority students and at-risk students as defined by free/reduced lunch. Table 4 below provides an interpretation for the size of the correlation coefficient for each of the cohorts.

Table 4

Correlation Coefficient Interpretation

Correlation Coefficient	Interpretation
.90 to 1.00 (-.90 to -1.00)	Very high positive (negative) correlation
.70 to .90 (-.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (-.30 to -.50)	Low positive (negative) correlation
.00 to .30 (.00 to -.30)	Little if any correlation

Note. Adapted from "Applied Statistics for the Behavioral Sciences" by Hinkle, Wiersma & Jurs, p. 109. Copyright 2003 by Houghton Mifflin.

A coefficient of determination (r^2) was determined for each of the cohorts and the sub-groups within each cohort. According to Hinkle et al. (2003), "The coefficient of determination indicates the proportion of the variance in one variable that can be associated with the variance in the other variable" (p. 110). The coefficient of determination was calculated by squaring the correlation coefficient. Thus, the researchers determined how much variance in KCCT test scores were determined by grades.

Regression Analysis

Finally, a regression analysis was used to measure the association of grades and the corresponding KCCT score between the two cohorts. According to Cronk (2010), "A simple linear regression allows the prediction of one variable from another" (p. 45). The

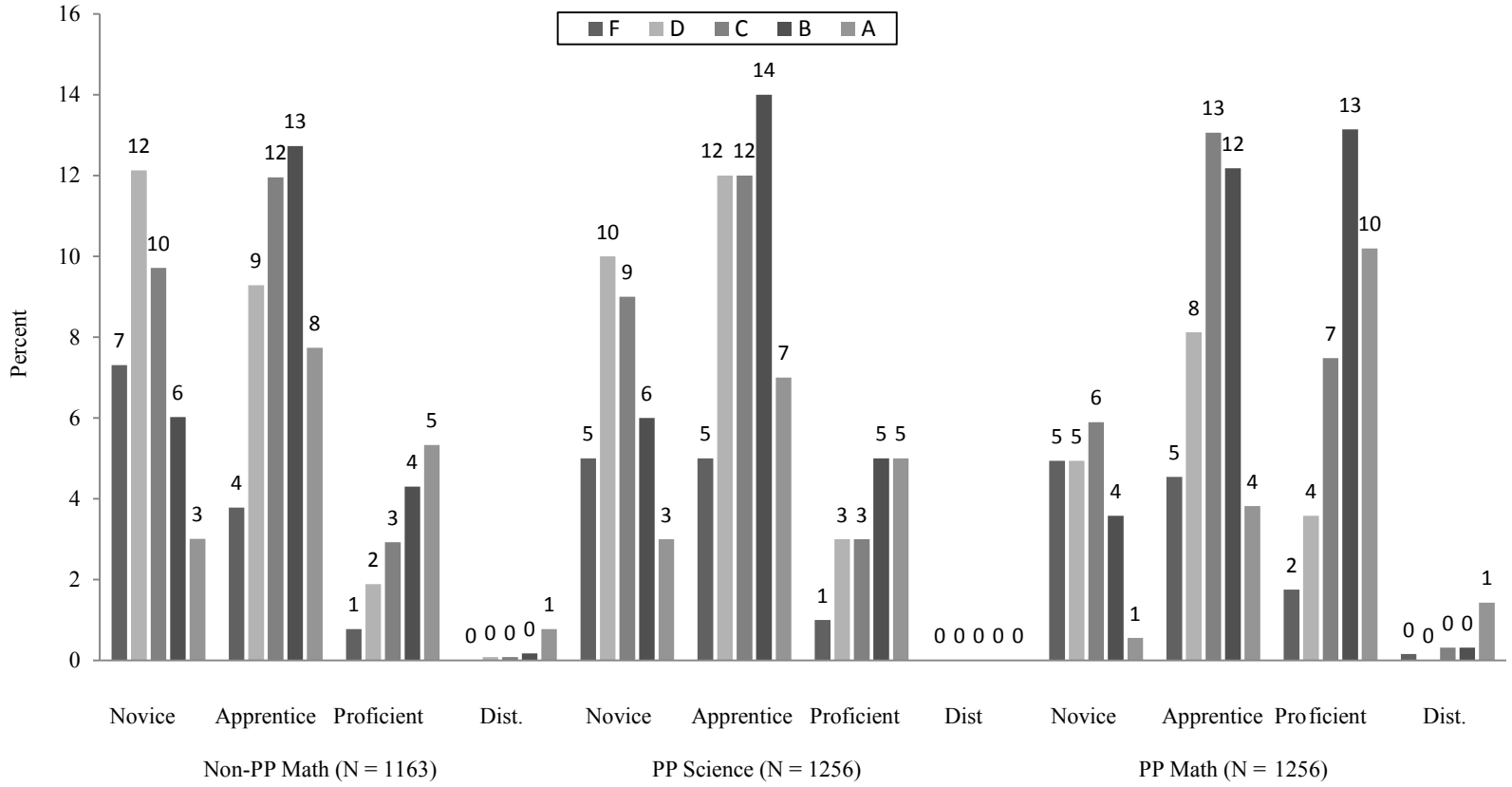
regression analysis allowed the researcher to determine the relationship between student grades and KCCT test scores in each of the cohorts. At-risk status (FRL), 8th grade math scale scores, and grades within the specific content course were used as variables to assess the relationship between grades and test scores. Other independent variables, such as at-risk status and prior academic achievement were used in the analysis to compare their impact on student achievement with course grades.

Results

Analysis of Student Grades and Test Scores

The researchers analyzed descriptive statistics of student grades and test scores to determine the percentage of students who received an above average grade in their math or science course (A/B) and also scored proficient or distinguished on the corresponding KCCT test. If students' grades are a valid indicator of student learning, then students scoring an A or a B in their content class should score proficient or above on the state accountability assessment. With the students who experienced traditional grading methods (the 2010 cohort), this assumption did not prove true. As Figure 1 demonstrates in the non-PP-Math cohort, 466students (40%) received an A or a B in their Algebra 2 class. Of these 466 students, only 123 (26%) scored a proficient or distinguished on the 2010 KCCT math assessment. Within the PP-Science group, in which students also experienced traditional grading, 514 students received an A or a B in their science class. Of these 514 students (41%), only 143 (28%) scored a proficient or distinguished on the 2011 KCCT science assessment. Success in the classroom as defined by grade did not translate into success on the KCCT assessment.

Figure 1. Grade distribution for Non-PP and PP classrooms. Dist. = Distinguished.



For students who experienced standards-based grading in PP-Math (the 2011 cohort), 568 (45%) received an A or a B in their Algebra II class. Of these 568 students, 315 (55.4%) scored proficient or distinguished on the 2011 KCCT math assessment. As a result of the treatment of PP, in which students were evaluated on standards-based grading methods, approximately twice as many students scored proficient or above on the state assessment as opposed to those students who were graded on traditional methods in non-PP-Math and PP-Science. Despite this increase in proficiency among students who experienced standards-based grading, educators should be concerned that 45% of the students in the PP-Math Cohort that achieved an A or a B in their Algebra II class still scored below proficient on the KCCT math assessment.

Relationship Between Grades and Student Achievement

Further data analysis was needed to assess the association between grades and achievement scores in both of the cohorts. A Pearson correlation coefficient was calculated for the relationship between participants' grades and KCCT test scores in each of the three groups. A low positive correlation was found in both PP-Math and non-PP-Math. Participants with higher grades tended to score higher on the KCCT test in both cohorts, although the magnitude was not as strong in non-PP-math as the cohort that experienced standards-based grading. Little if any correlation (Hinkle et al., 2003) was found in PP-Science when traditional grading was used as the correlation coefficients were the lowest of the three groups. Table 5 displays the results of the correlation data.

Further analysis revealed the Pearson correlation coefficients for each NCLB subgroups to determine the relationship between participants' grades and KCCT test

scores in each of the three cohort groups. A Pearson correlation coefficient was calculated for minority students in each of the three cohorts to determine the relationship between participants' grades and KCCT test scores. A low positive correlation was once again found in both PP-Math and non-PP-Math, while, little if any correlation was found with minority students in the PP-Science group in which traditional grading was used.

Table 5

Cohort Correlation Values

	PP-Math		Non-PP-Math		PP-Science	
Correlation	*r*	r^2	*r*	r^2	*r*	r^2
Overall	.45	.20	.38	.15	.26	.07
Minority	.39	.15	.36	.13	.23	.05
FRL	.44	.19	.39	.15	.27	.07

Note. FRL = free or reduced price eligible

Finally, a Pearson correlation coefficient was calculated for FRL students in each of the three cohorts to determine the relationship between participants' grades and KCCT test scores. A low positive correlation was also found for both the PP-Math and non-PP-Math cohorts, although, once again the magnitude was stronger for PP-Math students. Finally, little if any correlation was found in the PP-Science in FRL students in which traditional grading was used.

Correlational analyses found that students who experienced standards-based grading had both stronger correlations between grades and assessment scores and stronger coefficients of determination than students who experienced traditional grading. Despite an increased *r* and r^2 within PP-Math that used standards-based grading, only PP-Science fell below the low positive correlation as stated in Table 5. Students who experienced standards-based grading in PP-Math had a higher overall r^2 and a higher r^2 with minority and at-risk students as compared to the cohorts that used traditional grading

methods. Students who experienced standards-based grading in PP-Math also experienced traditional grading in their science class with PP-Science. The exact same students took the KCCT math assessment and also the KCCT science assessment. When students experienced standards-based grading, the correlation between grades and achievement scores was stronger with all students, minority students, and at-risk students then the correlation statistics when experiencing traditional grading. However, educators should again be concerned that in both cohorts, grades only had a low positive correlation with achievement scores. Even within the cohort that experienced standards-based grading, the association between grades and test scores was not as high as expected.

Analysis of Variance in Contrast of Grades

Further analysis of the data explored the mean scores between each grade for the three groups and the analysis of variance between grades within the groups. A one-way ANOVA compared the grades of participants in each cohort with the achievement score from the corresponding KCCT assessment. In PP-Math, non-PP-Math, and PP-Science, a significant difference was found between grades students received and their KCCT score. Table 6 displays the mean scores for each group based on grade.

Table 6

Mean Scale Scores per Grade

Grade	PP-Math[a]	Non-PP-Math[b]	PP-Science[c]
A	47.31	35.40	34.26
B	37.48	28.69	29.50
C	31.18	23.34	24.69
D	26.93	19.96	23.29
U	22.34	16.59	22.04

Note. [a]State mean = 37.00. [b]State mean = 36.00. [c]State mean = 36.00.

This analysis revealed that students in PP-Math who received an A in their math course ($M = 47.31$, $SD = 13.77$) scored higher on the KCCT assessment than students who received a B in their math course ($M = 37.48$, $SD = 15.11$). Post hoc tests revealed that students continued to score lower on the KCCT assessment based on their specific grade. The grade a student received was more strongly associated with their performance on the KCCT assessment for the PP-Math group than for the non-PP-Math and PP-Science groups. A larger variance between classroom grade and KCCT scores indicated that standards-based grades were more strongly associated with standardized test scores.

Significant differences between the standards-based grading cohort as compared to the traditional grading cohorts were also found in the analysis of variance among students who scored above average in their specific content class and their corresponding KCCT score. After an analysis and comparison of the mean scores of each of the cohorts based on the grade achieved in the specific content course, the researchers discovered further evidence to support the use of a standards-based grading approach. As seen in Table 6, students who achieved an A or a B in their math class when experiencing standards-based grading had a mean score of nearly 12 points higher than those students receiving an A or a B in the traditional grading cohorts. The difference in mean scale scores continued to decrease for students who achieved a C or a D until there was very little difference in mean scale score between students who failed the course whether they were evaluated on standards based or traditional grading methods.

Students who achieved an A or a B in their content course in PP-Math also scored above the state mean on the KCCT assessment. In non-PP-Math and PP-Science, even students who scored an A in their content course fell below the state mean on the

corresponding KCCT assessment. Therefore, students who achieved high grades in their math class with standards-based grading tended to score higher on the KCCT assessment than those receiving above average grades in the traditional grading cohort. Students who achieved low scores (D/U) in the class always tended to score low on the KCCT assessment, whether they were evaluated on a standards-based or traditional approach.

The data on standards-based grading was reinforced by analyzing the results of the ANOVA based on contrast in grades. Students who received an A in their math class with standards-based grading scored nearly 10 points higher on the KCCT assessment than students who received a B in their math course. The same students evaluated with traditional grading who received an A in their science courses only scored five points higher than students who received a B. The mean difference between students who scored an A in their content class as opposed to students who scored a B was much less in the traditional grading groups than standards-based grading groups. This trend continued throughout the analysis as students who were evaluated on standards-based grading scored higher on the KCCT assessment if they had a higher grade in the course.

The most disturbing data found that students who were evaluated on standards-based grading and received an A in their course scored nearly 25 points higher on the KCCT math assessment than those students who failed the course, yet students who were evaluated with traditional grading and achieved an A in the course scored only 12 points higher on the KCCT science assessment than those students who failed the course. As displayed in Table 7, the difference between students who received an A in the course using traditional grading practices and students who failed the course was not even one performance level on the KCCT assessment.

Table 7

KCCT Mean Scale Score Difference by Classroom Grade

Grade Contrast	PP-Math	Non PP-Math	PP-Science
A-B	9.83	6.71	4.78
A-C	16.13	12.07	9.57
A-D	20.36	15.44	10.98
A-U	24.97	18.82	12.23
B-C	6.31	5.36	4.79
B-D	10.55	8.73	6.20
B-U	15.15	12.11	7.45
C-D	4.24	3.38	1.40
C-U	8.84	6.75	2.66
D-U	4.59	3.73	1.25

Note. Values represent KCCT mean scale score differences for students receiving identified grades.

If KCCT scores are considered to be valid measures of student achievement, and if grades are to be a valid measure of student achievement, then differences in class grade should also equate to differences in achievement score. In the end, this analysis found that grades were more associated with KCCT scores in standards-based grading than with traditional grading methods.

Linear Regression Results

Table 8 shows the results of the linear regression.

Table 8

Summary of Multiple Regression for Non-PP and PP Cohorts

	Non-PP Math				Non-PP Science				PP Math			
Variable	B	*se*(B)	β	*t*	B	*se*(B)	β	*t*	B	*se*(B)	β	*t*
Constant	3.72	.94		3.96**	10.22	1.03		9.96**	9.85	.99		1.00**
SES	-1.55	.70	-.05	-2.20*	-1.77	.76	-.05	-2.34*	-.89	.73	-.02	-1.21
Achievement	.57	.02	.62	28.67**	.48	.02	.56	24.54**	.59	.02	.61	29.94**
Grade	2.72	.26	.22	10.46**	1.80	.27	.15	6.61**	3.38	.28	.25	12.23**

Note. $*p < .05$. $**p < .001$. B = unstandardized coefficient. β = standardized coefficient.

Prior academic achievement was the strongest predictor of KCCT achievement score in math or science for 11th grade students. Grades within the specific content course were also significant predictors of achievement on the corresponding KCCT assessment in the non-PP-Math cohort and with both PP-Math students and the same PP-Science students who did not receive the PP treatment. Although the grade within the course was significant in all three groups, it was a stronger predictor among students who experienced standards-based grading in PP than among students who experienced traditional grading in the non-PP groups. The standardized Beta coefficients were similar in the PP-Math cohort ($\beta = .25$, $p < .001$) and non-PP-Math cohort ($\beta = .22$, $p < .001$), and smaller for PP-Science students ($\beta = .15$, $p < .001$). The same students took both the math and science KCCT test during the same testing window. In PP-Math, the students experienced standards-based grading. In PP-Science, the same students experienced traditional grading methods. Through the regression analysis, the researchers found that math grades had a higher association with KCCT math scores than science grades had with KCCT science scores.

Discussion

With regard to the first research question, a stronger association clearly existed between course grades and standardized test scores among students who experienced standards-based grading as opposed to those students who experienced traditional grading methods. First, descriptive statistics found that more students who achieved an A or a B in their class scored proficient or above on state accountability testing when they experienced standards-based grading as opposed to traditional grading. Second, the Pearson correlation coefficient (r) and coefficients of determination (r^2) determined a

greater association. Third, the analysis of variance found that students who achieved higher grades in their math class also achieved higher scores on the KCCT assessment when they experienced standards-based grading. Finally, the regression analysis found that grades in a standards-based grading model accurately predicted a students' KCCT score at a higher rate than traditional science grades but not standards-based math grades.

In terms of the second research question, a stronger association was also found with at-risk students, although the association was slightly lower than the overall population. The analysis of correlation statistics found that the proportion in variance among FRL students in the relationship between grades and test scores was over twice as high for PP-Math students (19%) who experienced standards-based grading as compared to PP-Science students (7%) who experienced traditional grading. The same at-risk students had higher correlation statistics when evaluated on standards-based grading as opposed to traditional grading.

Finally, with regard to the third research question, minority students in the standards-based grading cohort had a stronger association between grades and test scores than minority students who experienced traditional grading, although the difference was the least of all the groups. The difference in variance between grades and test scores (r^2) of PP-Math students (15%) and non-PP-Math students (13%) was minimal. Despite this limited difference, the variance between grades and test scores was triple the amount for PP-Math students (15%) compared to PP-Science students (5%) who experienced traditional grading methods. Although this difference was substantial, minority students had the smallest difference between PP-Math cohort students and non-PP-Math cohort students. These findings support prior research that found that teachers grade minority

students less on content mastery and more on a hodgepodge of various factors (Brennan et al., 2001; Happonstall, 2010).

Conclusion

Based on the results of this research study, the use of standards-based grading with PP classrooms increased the association between grades and standardized test scores among students within the 11 high schools that implemented the program. Students who were more successful in the content course that used standards-based grading were much more likely to score proficient on the KCCT assessment than students evaluated on traditional grading practices. The most significant finding to refute traditional grading methods was that over 75% of students who received above average traditional grades in their specific content class scored below proficient on the corresponding KCCT assessment. Nearly twice as many students scored proficient when successful in their core content course when they were evaluated by standards-based grading. These findings argue for making standards-based grading approaches central to the education reform movement. As American education continues down the path of developing mandated state standards for teachers and schools, the evidence suggests that teachers should assess students on a standards-based grading model.

Although the association between grades and test scores was stronger with standards-based grading within this study, several limitations existed. First, the level of implementation of standards-based grading within each school and classroom was not explored. Teachers in each of the 11 schools implemented PP within their mathematics classrooms and this implementation required a certain level of fidelity with standards-based grading. The tenets of PP required that teachers "guarantee competency" of each

of their students on three key standards for each six weeks grading period. Without implementing a standards-based grading approach, teachers could not ensure that each student had met the three key standards. Schools and classrooms, however, could vary in their level of implementation with standards-based grading. This study did not take into account the level of fidelity of implementation with standards-based grading. Although science classes were used as a comparison group to measure the differences in a traditional grading approach and standards based grading approach with the same students, teachers most likely varied in their fidelity of implementation of PP and specifically standards-based grading. This research study did not account for these differences.

A second limitation of the study was the lack of correspondence between KCCT assessments and the tested course. The KCCT assessments in math and science assess content over three courses throughout a student's high school career. In mathematics, Algebra 1, Geometry, and Algebra 2 content are a part of the KCCT mathematics assessment. Within this study, the grades students received on a standards-based grading approach only evaluate students on Algebra 2 content. Therefore, a student could have successfully mastered the content in Algebra 2, but the student's standardized assessment score could have suffered because of a deficiency in a previous mathematics class. In order to truly assess the association between grades and test scores, the standardized assessment should only cover content taught in that specific course.

Implications for Policy, Practice, and Future Research

In an age of increased accountability and high-stakes testing, the implications of this research for practitioners become clear. Currently, schools are held accountable for

every student's proficiency in core content areas. Teachers are rewarded for their students' performance on state accountability assessments and likewise punished for the lack of performance on these accountability assessments (Podgursky & Springer, 2007). In order for both teachers and schools to ensure the learning of state standards by all students, a standards-based grading approach appears to offer a more valid method. This research demonstrated that students who performed above average in their class and were evaluated with a standards-based grading approach performed higher on state accountability assessments than those students who performed above average in their class with traditional grading. This finding supports the reasoning that the grades students receive in a core content class using standards-based grading reflect to a higher degree what the students can demonstrate on state proficiency assessments. As a result, grades in a standards-based assessment system are a more valid reflection of student learning (Allen, 2005).

Standards-based grading challenges many of the traditional school concepts that teachers and administrators have supported for decades (Brookhart, 1993; Cross & Frary, 1999; Guskey, 2007). When teachers evaluate students based on attainment of key standards, the onus of responsibility turns from the student to the teacher. No longer can practitioners fall back on the failure of students to submit homework, class work, or adhere to traditional policies within a classroom that have previously determined a students' grade. Now, when students do not pass a class, the teacher must face the reality that they may not have successfully taught the students the necessary skills to achieve the standard. In standards-based grading, teachers may not be able to blame student failure

on lack of compliance with the usual hodgepodge of factors that have resulted in a students' grade on traditional grading models.

Policymakers have worked to legislate curricula and teaching approaches most likely to result in a reduction in the achievement gap. This gap between minority and non-minority students and at-risk and not-at-risk students has been widely discussed. Very little, if any of the focus to reduce the achievement gap has centered around a change in grading practices. Within this study, evidence suggested that standardized test scores increased as a result of standards-based grading. Nearly twice as many students scored proficient on the mathematics assessment when they experienced standards-based grading in their Algebra 2 class. Therefore, policymakers should increase the scrutiny on traditional grading practices within schools and emphasize standards-based grading practices as a way to reduce the achievement gap in American education (Brennan et al., 2001; Happonstall, 2010).

Importantly, both practitioners and policymakers must grapple with ways to deal with the diligent student who is unable to master the key standards to attain a passing grade. With traditional grading systems, a student who is compliant with a teacher's policies and requests, completes all assigned tasks in a timely manner, and has a good attendance and behavior record almost always passes a core content class and is provided by his or her teachers with signals (grades) that they are performing at competent levels and are likely to pass the state's tests. With standards-based grading, these hodgepodge factors will have little influence on a student's grade. If teachers are going to grade on standard-attainment, then the student who does not attain the standard must be given a accurate academic feedback, e.g., a failing grade. With the emphasis on having all

students college ready based on benchmark ACT scores, schools should consider using standards-based grading approaches to ensure that students are truly college ready. At the same time, schools are expected to graduate all students and decrease retention rates. Policymakers and practitioners should explore methods through which to measure students on attainment of key academic standards, while still providing necessary safety nets for students who are unable to achieve these standards.

Finally, researchers should turn their attention towards the effectiveness of standards-based grading practices in schools. Most prior research has centered on overall grading practices, and little empirical research exists to support a movement towards standards-based grading. Researchers can build on the data within this study to establish a strong empirical research base for the widespread implementation of standards-based grading. As with the findings from this study, researchers can continue to make the case that high grades in a content class that use standards-based grading have a higher correlation with high standardized test scores than high grades in a content class using traditional grading practices. Future research can use new end-of-course assessments to measure the association between grades and test scores within this new format. This research should also focus on the impact of standards-based grading on the attainment of key standards by minority students and at-risk students. As the standards-based grading movement continues to grow in secondary schools, researchers should explore the potential reduction in the achievement gap as a result of new grading practices.

Researchers should also measure the level of implementation with standards-based grading and the reactions to this movement from school stakeholders that include teachers, students and parents. This implementation will require researchers to develop

qualitative and mixed method research on standards-based grading that include observation and interview data with these critical stakeholders to assess levels of implementation of grading practices.

PROJECT PROFICIENCY: ASSESSING THE INDEPENDENT EFFECTS OF URBAN REFORM ON STUDENT ACHIEVEMENT

In the new landscape of American education, scaling up dramatic and sustainable instructional improvements in historically low-achieving schools, and doing so quickly are key challenges for educators and policymakers. With the advent of No Child Left Behind (NCLB), Race to the Top (RTTT), and School Improvement Grants (SIG), schools' and districts' leaders are mandated to search for quick, effective, and sustainable changes in instruction, assessment, and intervention to improve student results and avoid state and federal sanctions. Unfortunately, a wide body of research over the past half-century (Aiken, 1942; Austin, 1979; Crandall et al., 1982; McLaughlin, 1990; Slavin et al., 1996; Stringfield, Reynolds, & Schaffer, 2008; Tucker, 2011) has found shortfalls in taking reform to scale. When school or district level reform has occurred, the time needed to root innovation in the instructional culture of a school has typically not been sufficient to see gains that meet federal and state guidelines.

The purpose of this study is to examine the impact of Project Proficiency (PP), an instructional reform adopted in one urban school district to enhance teaching and learning in math and English courses. To achieve this purpose, the author uses a nonequivalent control group design and Hierarchical Linear Modeling (HLM) to address the following two research questions:

1. Does Project Proficiency implementation increase academic performance on state mathematics core content assessments?
2. Does Project Proficiency implementation reduce the variance of classroom compositional effects on state mathematics core content assessments?

Project Proficiency: Ensuring Learning in One Urban School District

With 10 of 21 comprehensive high schools in Louisville, Kentucky designated as "Persistently Low Achieving" (PLA) by state officials, district leaders in Jefferson County Public Schools (JCPS) confronted the need to increase dramatically the number of students scoring proficient or higher on the Kentucky Core Content Test (KCCT) (Kentucky Department of Education, 2008). JCPS leaders created PP, an initiative designed to "balance instructors' commitment to teaching with their ownership of student learning" (Jefferson County Public Schools, 2011a, p. 3). PP was created in order to "guarantee competency" through the implementation of four practices: (1) focusing on three key standard categories in reading and mathematics classes during each grading period, (2) creating formative tasks (Stiggins, Arter, Chappuis, & Chappuis, 2004) through which students demonstrate competency for each standard, (3) guiding each student to a level of competency in each standard prior to the end of each six-week grading period, and (4) creating fail-safe assessments that enable students to recover missed content (Jefferson County Public Schools, 2011a). Initially during the 2010-2011 school year, these practices were adopted only in Algebra 2 and English 2 courses.

District leaders encouraged schools participating in PP to create their own operationalizations of the four key elements of the design. To assist teachers, Professional Learning Communities (PLC) and principals, the district refined its online Classroom Assessment System and Community Access Dashboard for Education (CASCADE) to track diagnostic, formative and summative assessment results. Required

implementation of PP provided schools with autonomy to design, engineer, and develop their work in a manner that reflected the unique characteristics of a school, while holding a few key concepts universal to the design (Bryk, 2009; Stringfield et al., 2008).

The guiding premise behind PP is that all students must demonstrate a level of competency through diagnostic assessments and formative work before taking an end-of-unit proficiency examination. As opposed to traditional instructional programs in which a teacher assigns a grade and continues with the scope and sequence of the course and creates what Stiggins (2007) calls "losing streaks" for students not mastering content, district and school leaders charged PP teachers with creating new and innovative ways to ensure learning of three key standards for each unit of study. PLCs (Abbott & Fisher, 2011; Allen & Blythe, 2004; DuFour, DuFour, Eaker, & Karhanek, 2004; DuFour & Eaker, 1998; McDonald, Mohr, Dichter, & McDonald, 2007) provided a mechanism for developing teachers' collective efficacy. Teams of teachers met to plan for formative assessments, examine formative work, and adjust practices by reviewing student data.

The actions by school and district leaders to operationalize PP mirror the "strategic functions" identified by Childress, Elmore, Grossman, and Akinola (2006) to support reform efforts in urban school districts. First, leaders created a common instructional framework that served as the foundation for the design. Teachers received a narrow and focused curriculum and were challenged to balance formative and summative assessments. These conditions encouraged teachers to share accountability for student achievement with both the student and the teacher's PLC. Second, school and district leaders created structures to support PP implementation. For example, common planning time was scheduled to provide teachers with an opportunity to collaborate during the

school day, and master schedules were altered to place PP courses during the same period to allow teachers to regroup students needing additional support across classes. These structures enabled English and Algebra teachers to work in a collaborative instructional environment where all teachers owned the learning results of their students. District leaders fostered collaboration between PP principals and teachers to share best practices and support improvements in all classrooms (Fullan, 2011). Finally, the district provided coordinated resources to PP schools in the form of professional development activities, a data management system tailored to PP implementation, and curricular materials that ensured instructional coherence within and between schools implementing PP.

The Influence of Quality Teaching

District and school leaders implemented PP structures to make dramatic and lasting changes in both teaching practice and student performance, as a strong link exists between the two (Atherton, 2011; Brophy, 1988; Brophy & Good, 1986; Darling-Hammond & Youngs, 2002; Goldhaber & Brewer, 1997; Hattie, 2003, 2009; Marzano, Pickering, & Pollock, 2001; Muñoz & Chang, 2007). Students, regardless of family or academic background, benefit from high quality teaching. In their value-added analysis of teaching effects in Chicago high schools, Aaronson, Barrow, and Sander (2007), noted that a one standard deviation improvement in teacher quality equated to a 22% increase in mathematics achievement and that African-American students benefited most from a higher-quality teacher. From his study of 4600 elementary students and 307 teachers, Stronge (2010) found that after controlling for student-level factors, such as prior academic achievement and background characteristics, achievement results at the end of one year of instruction ranged from approximately -2 to +2 standard deviations from

classroom to classroom. Stronge (2010), summarizing his findings in a manner consistent with decades of teacher effectiveness studies, concluded that, "the quality of a teacher that a student happens to be assigned will play an extraordinary role in the student's academic success, at least the time the student is under the teacher's tutelage, and perhaps beyond" (p. 11).

In their review of research on teacher and classroom effects on achievement, Odden, Borman, and Fermanich (2004) noted that a school's impact on student learning can be attributed to the cumulative effect of its teachers, noting that the experience and education level of the faculty in addition to the quality of professional collaboration present in a school enhances instructional ability and improves student achievement. High-quality instructional practices also strengthen teaching and subsequent learning. In a meta-analysis of over 500,000 studies on the influences of student achievement, Hattie (2003) noted that teacher qualities such as student feedback, direct instruction, and remediation have the largest effect sizes. Through the use of HLM, Hattie (2003) also noted that approximately 30% of the variance in student achievement can be attributed to teachers and stated that, "excellence in teaching is the single most powerful influence in achievement" (p. 4).

The Confounding Effect of Socioeconomic Status on Student Achievement

A school's socioeconomic makeup and the effect that large concentrations of underperforming students have on all facets of the school are important considerations for school and district leaders when creating instructional programming and evaluating its effectiveness. The average socioeconomic status (SES) of a school's student body, defined as the percentage of students attending the school that qualify for free or reduced

price lunches, impacts student achievement (Berends & Peñdoza, 2010; Coleman et al., 1966; Hochbein & Duke, 2011; Jencks, 1972; Mickelson, 2010; Palardy, 2008; Perry & McConney, 2010; Rumberger & Palardy, 2005; Teddlie, Stringfield, Wimpleberg, & Kirby, 1987). Schools with large compositions of students with risk factors typically have difficulty meeting academic goals, recruiting and retaining high-quality teachers, and sustaining innovation (Darling-Hammond, 1997; Payne, 2008; Popham, 1997). Borman and Dowling (2010) used HLM to re-examine the *Equality of Education Opportunity* (EEO) data used in the seminal Coleman et al. (1966) Report. Through HLM modeling, the authors determined that after controlling for family background and school, teacher, and peer effects, the racial and socioeconomic composition of a school impacts student achievement. Borman and Dowling (2010) found that school-level African-American enrollment and SES effects are one and three-fourths times more powerful than an individual's race or socioeconomic status in predicting student achievement. Whereas Coleman et al. (1966) reported that school composition and contextual factors (described below) explained only 4% in variability in student achievement beyond the contribution made by individual students' background, Borman and Dowling's (2010) analysis demonstrated that school composition alone explains nearly 25% of variability in achievement.

Borman and Dowling's (2010) examination of contextual factors such as curriculum, resources, and academic and nonacademic tracking practices found that teachers' biased perceptions of students from low socioeconomic backgrounds significantly widened the black-white achievement gap. Furthermore, the researchers found that 40% of variability in achievement existed between schools and provided

evidence that schools do in fact influence student achievement. Though Borman and Dowling's (2010) findings mirrored those of Coleman et al. (1966) regarding the effect of student background on achievement, their re-analysis demonstrated the greater "compositional" effect that schools have in determining student achievement. Borman and Dowling (2010) summarized their study by saying that, "these findings reveal that school contextual effects dwarf the effects of family background" (p. 1239). Their insight is important to educators and policymakers who seek to evaluate reform efforts and their potential scalability. The quality of instruction, curriculum, and assessment programs within a school has important implications for educators wishing to implement instructional programs designed to assist students from disadvantaged backgrounds. Though school SES was found to be a confounding factor in student achievement, changes in teaching and classroom practices were also found to yield gains in student achievement (Ballou, Sanders, & Wright, 2004; Brown-Jeffy, 2009).

Using HLM to Examine Student Achievement

To obtain a balanced and accurate picture of the educational landscape, researchers must examine the influence of teaching and compositional effects on student achievement and their relationship with one another. As previously described in studies by Borman and Dowling (2010) and Hattie's (2003) meta-analyses, researchers have utilized sophisticated methodologies for examining instructional innovations like PP. HLM is widely regarded as an effective statistical method to study the effects of instructional practices at the classroom, school, and district levels (Raudenbush & Bryk, 1986, 2002). Unlike traditional OLS regression modeling techniques, HLM enables researchers to address shortcomings with single-level statistical modeling by nesting data

at the individual, classroom, and school level and to determine the impact that covariates have on student and classroom growth trajectories. By clustering data in this manner, researchers can more accurately estimate individual effects, model cross-level effects, and partition variance to estimate covariate relationships (Raudenbush & Bryk, 2002). Researchers have used HLM to determine the impact that the following factors have on student achievement: SES (Palardy, 2008; Rumberger & Palardy, 2005), prior achievement (Lee & Bryk, 1989; Raudenbush, 2004), gender (Stronge, 2010), ethnicity (Berends & Peñdoza, 2010) school composition (Brown-Jeffy, 2009; Hanushek, Kain, & Rivkin, 2009; Mickelson, 2010; Willms, 2010), school size (Xiang & Hauser, 2010), and school achievement (Tewke et al., 2004). These authors reported that students' prior academic achievement and classroom and school SES are powerful predictors of student performance (Borman & Dowling, 2010). In this study, HLM was used to determine the effect that PP had on student achievement after controlling for student-level and classroom-level characteristics, since the literature suggests that prior achievement and classroom compositional effects significantly impact student achievement.

Method

Population and Study Participants

The researchers drew participants from 11 high schools that implemented PP in JCPS during the 2010-2011 school year. The school district is the 26th largest in the US, enrolls 100,474 students across 150 schools, and serves one-seventh of all students in Kentucky. The demographic composition of the district's students is as follows: 51% White, 37% African-American, and 12% other. Approximately 62% of JCPS' students qualify for free or reduced-price lunches. Eight of the 11 schools in the study were

identified as Persistently Low Achieving (PLA) for not making Adequate Yearly Progress (AYP) in combined state reading and mathematics proficiency measures and were subject to state and federal sanctions (Persistently Low-Achieving School, 2010; School Improvement Fund, 2010). Four of the schools were in year one of state-mandated consequences and the remaining four were subject to consequences at the end of the 2010-11 school year. Data from before PP implementation (the 2009-2010 school year) and after initial PP implementation (the 2010-2011 school year) were used in the current analysis. Table 9 provides demographic information for the schools being studied.

Table 9

Demographic Characteristics within Cohorts (N = 2451)

	Non-PP Cohort		PP Cohort	
Characteristic	n	%	n	%
Gender				
Male	585	50.3	629	48.8
Female	578	49.7	659	51.2
Race/ethnicity				
Caucasian/White	509	43.8	567	44.0
African-American	586	50.4	639	49.6
Hispanic	52	4.5	58	4.5
Other Minority	16	1.4	24	1.9
Free/reduced lunch	822	70.7	946	73.4
Total	1163	100	1288	100

In order to nest classrooms for HLM analysis, school-level master schedules were analyzed to identify Algebra 2 classrooms in the population schools. The 2009-10 group sample contains 1163 students in 104 mathematics classrooms. The 2010-11 group sample contains 1288 students in 110 mathematics classrooms. Table 10 identifies classroom characteristics.

Table 10

School and Classroom Characteristics

	Non-PP Cohort (2009-2010)			PP Cohort (2010-2011)		
School	Students	Classrooms	% FRL	Students	Classrooms	% FRL
1	155	11	82.7	169	9	82.0
2	122	11	69.8	123	10	74.4
3	102	10	70.2	115	12	70.4
4	142	13	52.2	177	13	54.0
5	114	10	58.5	124	12	57.4
6	73	5	65.5	88	7	67.9
7	35	3	85.6	66	6	85.9
8	154	11	67.9	143	11	67.5
9	125	13	76.6	116	10	75.5
10	62	7	66.7	90	11	71.8
11	79	9	82.1	77	8	80.9

Note. % FRL denotes percentage of students in school free or reduced price eligible.

Research Design

The study used a quasi-experimental nonequivalent control group design (Rossi, Lipsey, & Freeman, 1999; Shadish, Cook, & Campbell, 2002; Trochim & Donnelly, 2008) to analyze the effects of PP on grade 11 mathematics during the 2010-11 school year. In their explanation of quasi-experimental designs, Rossi et al. (1999) called for researchers to exercise a strong degree of care to ensure matching between comparison groups. The authors recommend that the evaluator first identify the intervention group and then construct a control group by selecting individuals not exposed to the treatment, but similar in characteristics of the treatment group. Two control groupings were used to compare the effects that PP had on mathematics achievement in the targeted schools. First, students assessed during the 2009-10 school year that did not experience PP in mathematics served as a first control group. Second, the researchers analyzed performance by PP students on an assessment other than mathematics. Use of these control groups enabled the researchers to compare achievement gains between years and

to determine the effect that PP had on student achievement in mathematics as compared to social studies with the same students during the same year.

Measures

This study used HLM to examine relationships within and between classrooms and their combined impact on grade 11 mathematics. To accomplish this, the researchers analyzed both student- and classroom-level independent variables. Level 1 independent variables are student characteristics used to examine relationships at the student level and Level 2 variables are independent variables used to examine relationships at the classroom level.

In this study, two dependent variables from the KCCT were used: Grade 11 Mathematics Mean Scale Score (11Math) and Grade 11 Social Studies Mean Scale Score (11SocStu). 11Math served as the primary dependent variable and 11SocStu served as an additional dependent variable in the study's nonequivalent control group design. All student and classroom data in this study were obtained from the Jefferson County Public Schools Data Warehouse. Mean scale scores (MSS), ranging from 0-80, came from KCCT administrations to students in grade 11. The KCCT Test Administration Guide (Kentucky Department of Education, 2008) identifies Cronbach's alpha measures to report internal consistency and reports an average measure of .89 and .84 for the 6 versions of the mathematics and social studies tests, respectively.

Students included in the study completed an Algebra 2 course during the 2009-10 or 2010-11 school years and have grade 11 KCCT results in mathematics and social studies during the same year. For purposes of this study, Algebra 2 is defined as a course in the Kentucky Program of Studies (Minimum Requirements for High School

Graduation, 2011) that meets the Kentucky Algebra 2 graduation requirement. These courses include Algebra 2, Algebra 2 Honors, and Algebra 2 Advanced, which contained mainstreamed students who qualify for special education services. Students who did not complete an Algebra 2 course or have a KCCT combination in mathematics and social studies during the same academic year were excluded from the sample.

As described in the literature review, student SES is a confounding factor in assessing the effect of reforms on academic achievement. Student SES (StudentSES) served as a first Level 1 predictor variable measured by using student eligibility for free or reduced-price lunch (Perry & McConney, 2010; Raudenbush & Willms, 1995; Rivkin, Hanushek, & Kain, 2005). Prior student achievement served as a second Level 1 independent variable (Lee & Bryk, 1989; Raudenbush, 2004; Raudenbush & Willms, 1995). Grade 8 KCCT mathematics (8Math) was a predictor for Level 1 analysis and grade 8 KCCT social studies (8SocStu) served as an independent variable for the control group. 8Math and 8SocStu MSS are reported on a 0-80 scale (Kentucky Department of Education, 2008).

Procedure and Models

Individual classroom information was nested for Level 2 analysis. Participation in PP was the primary Level 2 variable and classroom socioeconomic status (ClassSES) was used as an additional independent variable in the HLM analysis. ClassSES was measured by calculating the percentage of students in each classroom qualifying for free or reduced price lunches. Classroom SES was selected to control for differences in classrooms, as higher concentrations of low SES students typically appear in comprehensive and special education classrooms (Brown-Jeffy, 2009) and variability in

achievement has been connected to classroom tracking practices (Borman & Dowling, 2010).

To model the relationships between PP implementation and student performance, the researchers used HLM 7 software (Raudenbush, Bryk, & Congdon, 2011) to account for the hierarchical nature of student data nested within classrooms. The HLM analysis consisted of four parts: (1) a one-way ANOVA with random effects (unconditional model), (2) a regression with means-as-outcomes model controlling for classroom SES and PP implementation at the classroom level (Level 2), (3) a random-coefficient regression model controlling for student SES and prior academic achievement at the student level (Level 1), and (4) an intercepts- and slopes-as-outcomes model controlling for student SES and prior academic achievement at the student level (Level 1) and classroom SES and PP implementation at the classroom level (Level 2).

One-way ANOVA with random effects (unconditional model). The one-way ANOVA with random effects was used to determine how much variation in 11Math lies within (σ^2) and between (τ_{00}) classrooms and to examine the reliability of the Level 1 sample mean to estimate the mean of the entire population of schools studied (Raudenbush & Bryk, 2002). The equation below represents the one-way ANOVA with random effects model:

$$\text{Level 1: } 11Math_{ij} = \beta_{0j} + r_{ij}$$
$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}$$

The one-way ANOVA with random effects enabled the researcher to partition the total variation in the outcome variable. The intraclass correlation (ICC) was used to measure the proportion of variance in 11Math that is between classrooms and determine if further

HLM analysis was appropriate (Kreft & De Leeuw, 1998; Raudenbush & Bryk, 2002). The equation below represents the ICC calculation:

$$\hat{\rho} = \tau_{00}/(\sigma^2 + \tau_{00})$$

The variance components for random effects in this model were reported as the variance of the school true means around the grand mean (u_{0j}) and the Level 1 effect (r_{ij}) to list maximum likelihood (ML) estimates at the student level. The weighted least squares measurement (WLS) ($\hat{\gamma}_{00}$) was used to estimate the fixed effects of grand-mean math achievement. Deviance estimates were calculated to measure model fit (Kreft & De Leeuw, 1998).

Regression with means-as-outcomes model. The regression with means-as-outcomes model determined the effect that ClassSES and PP had on mean mathematics achievement. For the Level 2 analysis, grand mean centering was used to analyze ClassSES by subtracting individual classroom SES from the mean SES of the classrooms in the study. For this study, the researchers created dummy values for PP implementation (no Project Proficiency = "0," Project Proficiency = "1") for each classroom at Level 2. The equation below represents the regression with means-as-outcomes model:

Level 1: $11Math_{ij} = \beta_{0j} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(ClassSES_j - \overline{ClassSES.}) + \gamma_{02}(PP_j) + u_{0j}$

The regression with means-as-outcomes model enabled the researchers to estimate fixed effects and examine the variance of random effects. Analysis of fixed effects, as measured by the ClassSES (γ_{01}) and PP coefficients (γ_{02}), yielded an estimate of the strength of association between ClassSES and PP on 11Math achievement. By using the results obtained from the regression with means-as-outcomes model, the researchers

developed a proportion of variance explained index to determine if reductions in variance between schools were caused by the addition of Level 2 predictors. The proportion of variance explained index is as follows:

$$\hat{p} = \frac{\hat{\tau}_{00}(\text{random ANOVA}) - \hat{\tau}_{00}(\text{means-as-outcomes})}{\hat{\tau}_{00}(\text{random ANOVA})}$$

A conditional intraclass correlation (CICC) was created from the values obtained in the regression with means-as-outcomes model to determine if reductions in variance between pairs of scores within classrooms occurred after removing the effect of ClassSES to measure the degree of dependence among scores within classrooms that are of the same ClassSES (Raudenbush & Bryk, 2002). The equation below represents the CICC calculation:

$$\hat{\rho} = \tau_{00} / (\sigma^2 + \tau_{00})$$

Random-coefficient regression model. The random-coefficient model determined how much of classroom distribution of mathematics achievement was characterized by StudentSES and 8Math. For the Level 1 analysis, the researchers introduced dummy values for StudentSES (paid lunch = "0," free/reduced lunch eligible = "1"). The growth model for the random-coefficient analysis is as follows:

$$\begin{aligned} \text{Level 1:} \quad & 11Math_{ij} = \beta_{0j} + \beta_{1j}(StudentSES_{ij}) + \beta_{2j}(8Math_{ij}) + r_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + u_{0j} \\ & \beta_{1j} = \gamma_{10} + u_{1j} \\ & \beta_{2j} = \gamma_{20} + u_{2j} \end{aligned}$$

The variance components for random effects in this model were reported as the variance of the school true means around the grand mean (u_{0j}), increments to the slope by adding StudentSES (u_{1j}) and 8Math(u_{2j}), and the Level 1 effect (r_{ij}) to list ML estimates

at the student level. The generalized least square (GLS) measurement ($\hat{\gamma}_{00}$) was used to estimate the fixed effects of grand-mean math achievement and the addition of average Student SES ($\hat{\gamma}_{10}$) and 8Math ($\hat{\gamma}_{20}$) slopes across schools. The significance of the association between the Level 1 variables on the outcome variable was reported.

By using the results obtained from the random-coefficient model, the researchers developed a proportion of variance explained index to determine reductions in variance that occurred with the addition of the two Level 1 predictors. The proportion of variance explained index is as follows:

$$\hat{p} = \frac{[\hat{\sigma}^2 \text{(random ANOVA)} - \hat{\sigma}^2\text{(StudentSES)}]+[\hat{\sigma}^2\text{(random ANOVA)} - \hat{\sigma}^2\text{(8Math)}]}{\hat{\sigma}^2\text{(random ANOVA)}}$$

Intercepts- and slopes-as-outcomes model. The intercepts- and slopes-as-outcomes model was used to determine variability in 11Math across classrooms in schools implementing PP and identify how much variation in the slopes was explained by using ClassSES and PP as predictors. In the HLM model, the Level 1 equation remained the same as in the random-coefficient model; however, the Level 2 equation was expanded to incorporate ClassSES and PP. Similar to the means-as-outcomes model, grand mean centering was used in the Level 2 equation to analyze ClassSES by subtracting individual classroom SES from the mean SES of the classrooms in the study. Grand mean centering was used to remove high correlations between first- and second-level variables and cross-level interactions (Raudenbush & Bryk, 2002). The intercepts- and slopes-as-outcomes model is as follows:

$$\text{Level 1: } 11Math_{ij} = \beta_{0j} + \beta_{1j}(StudentSES_{ij}) + \beta_{2j}(8Math_{ij}) + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Class}SES_j - \overline{ClassSES.}) + \gamma_{02}(PP_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(ClassSES_j - \overline{ClassSES.}) + \gamma_{12}(PP_j) + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(ClassSES_j - \overline{ClassSES.}) + \gamma_{22}(PP_j) + u_{2j}$$

The variance components for random effects in this model were reported as the variance of the school true means around the grand mean (u_{0j}), increments to the slope by adding StudentSES (u_{1j}) and 8Math(u_{2j}), and the Level 1 effect (r_{ij}) to list ML estimates at the student level. The GLS measurement ($\hat{\gamma}_{00}$) was used to estimate the fixed effects of grand-mean math achievement and the addition of average StudentSES ($\hat{\gamma}_{10}$) and 8Math ($\hat{\gamma}_{20}$) slopes across schools at Level 1 in addition to ClassSES ($\hat{\gamma}_{21}$) and PP ($\hat{\gamma}_{22}$) at Level 2. The significance of the cross-level interactions between the Level 1 and Level 2 variables on the outcome variable was reported.

In addition to calculating a new proportion of variance explained (previously described in the random-coefficient model), the researchers developed a reduction in variance or variance-explained statistic at Level 2 for each of the random coefficients (intercepts and slopes) from Level 1. The proportion of variance explained index is as follows:

$$\hat{p} = \frac{\hat{\tau}_{qq}\,(\text{random-coefficient model}) - \hat{\tau}_{qq}\,(\text{intercepts- and slopes-as-outcomes})}{\hat{\tau}_{qq}\,(\text{random-coefficient model})}$$

By comparing the proportion of variance explained calculations in the means-as-outcomes, random-coefficient, and intercepts- and slopes-as-outcomes regression models, the researchers determined PP implementation produces reductions in variance when controlling for Level 1 and Level 2 independent variables.

Control Groups

For this study, two levels of controls were used to compare the effects that PP had on student mathematics achievement. Controlling for PP classrooms as a Level 2 variable in the regression with means-as-outcomes and intercepts- and slopes-as-outcomes models served as one method of determining the effect that PP implementation had on student achievement during the first year of implementation. The second involved conducting an additional HLM analysis using 11SocStu as the outcome variable. Since PP was not implemented in social studies classes during the treatment year, using 11SocStu as an outcome variable and comparing it to 11Math enabled the researchers to control for the internal validity threat of selection (Trochim & Donnelly, 2008). As for 8Math, 8SocStu served as the Level 1 predictor of student achievement. Use of these controls allowed the researchers to examine the independent effects of PP in the absence of the ability to select study group participants randomly as in traditional experimental designs.

Results

HLM Models- Math as Outcome

One-way ANOVA with random effects (unconditional model). The one-way ANOVA with random effects was used to examine initial variance relationships and determine if further HLM analysis was necessary. Tables 11 and 12 present the one-way ANOVA results. With regard to fixed effects in the one-way ANOVA, the WLS estimate for grand-mean 11Math achievement was 28.50, had a standard error of .73, and yielded a 95% confidence interval of 28.50 ± 1.96 (.73) (Appendix 1). To gauge the magnitude of variation among classrooms in mean 11Math achievement levels, the plausible range

of values was (9.90, 47.10) (Raudenbush & Bryk, 2002). A chi-square test was performed to examine the presence of variation among classrooms in 11Math achievement. Significant variation at the $p < .01$ level was found among classrooms in 11Math achievement. These findings indicated a substantial range in average achievement levels among classrooms in this sample of data and supports the need for additional HLM analysis. The within-school variance across classrooms was estimated at 203.37 and the between-classroom variance was 90.02 (Table 12). The ICC or proportion of total variance between classrooms was (90.02/(90.02 + 203.37) = .31. This suggested that 31% of variance in 11Math scores existed between classrooms and verified the need for further HLM analysis to explain within- and between-classroom variance relationships (Kreft & De Leeuw, 1998; Raudenbush & Bryk, 2002). The deviance statistic for the baseline model is reported in Appendix 1.

Means-as-outcomes regression model. For the means-as-outcomes regression model, the Level 1 equation remained the same, as 11Math scores were viewed as varying around classroom means. The Level 2 equation was expanded to examine how each classroom's mean was predicted by ClassSES and PP. With regard to fixed effects, a significant negative association was found between ClassSES ($\gamma_{01} = -13.30$, $t = -3.31$, $p < .01$) and 11Math, while a highly significant positive association was found between PP ($\gamma_{02} = 8.58$, $t = 6.80$, $p < .001$) and 11Math (Table 11). The residual variance of 61.38 between schools in the means-as-outcomes model was substantially smaller than the original value of 90.02 estimated in the one-way ANOVA model (Table 12). A range of plausible values for 11Math in this model was (8.66, 39.36). Though these were a wide range of plausible values, they were smaller than the range of values when ClassSES and

PP were not held constant, (9.90, 47.10). Chi-square tests determined that 11Math achievement means varied significantly ($p < .001$) after controlling for the effects of ClassSES and PP (Appendix 1).

Table 11

Fixed Effect HLM Models: Grade 11 Math as Outcome

Fixed Effect	Unconditional Model	Model 1	Model 2	Model 3
Classroom Means				
Intercept, γ_{00}	28.50***	24.01***	12.08***	7.59***
ClassSES, γ_{01}		-13.30**		.54
PP, γ_{02}		8.58***		8.00***
StudentSES-achievement slopes				
Intercept, γ_{10}			-.54	-.59
ClassSES, γ_{11}				-3.04
PP, γ_{12}				.20
8Math-achievement slopes				
Intercept, γ_{20}			.60***	.60***
ClassSES, γ_{21}				-.16*
PP, γ_{22}				.03

Note. *$p < .05$. **$p < .01$. ***$p < .001$. Model 1: Means-as-Outcomes. Model 2: Random-Coefficient. Model 3: Intercepts- and Slopes-as-Outcomes.

By comparing the between-class estimates across the one-way ANOVA and regression with means-as-outcomes models, the proportion reduction in variance explained was .32 [(90.02 – 61.38)/90.02 = .32] (Table 13). That is, 32% of the true between-classroom variance in 11Math was accounted for by ClassSES and PP. After removing the effect of ClassSES and PP, the correlation between pairs of scores in the same classroom, which had been .31, was now reduced to .23 by calculating the CICC: 61.38/(61.38 + 203.37) = .23. With the information obtained from the random-coefficient regression model, we can determine that classroom SES had a negative impact on student achievement, PP positively impacted mathematics achievement, and the

combination of the two variables reduced variation in mathematics scores between classrooms by approximately 25% over the unconditional model.

Table 12

Random Effect HLM Models: Grade 11 Math as Outcome

Random Effect	Unconditional Model	Model 1	Model 2	Model 3
Classroom Mean, u_{0j}	90.02***	61.38***	32.23*	20.39**
StudentSES-achievement slope, u_{1j}			.45	.50
8Math-achievement slope, u_{2j}			.00	.00
Level-1 effect, r_{ij}	203.37	204.80	122.21	123.33
Intraclass Correlation	.31	.23	.18	.14
Proportion Reduction in Variance Explained		.32	.40	.37

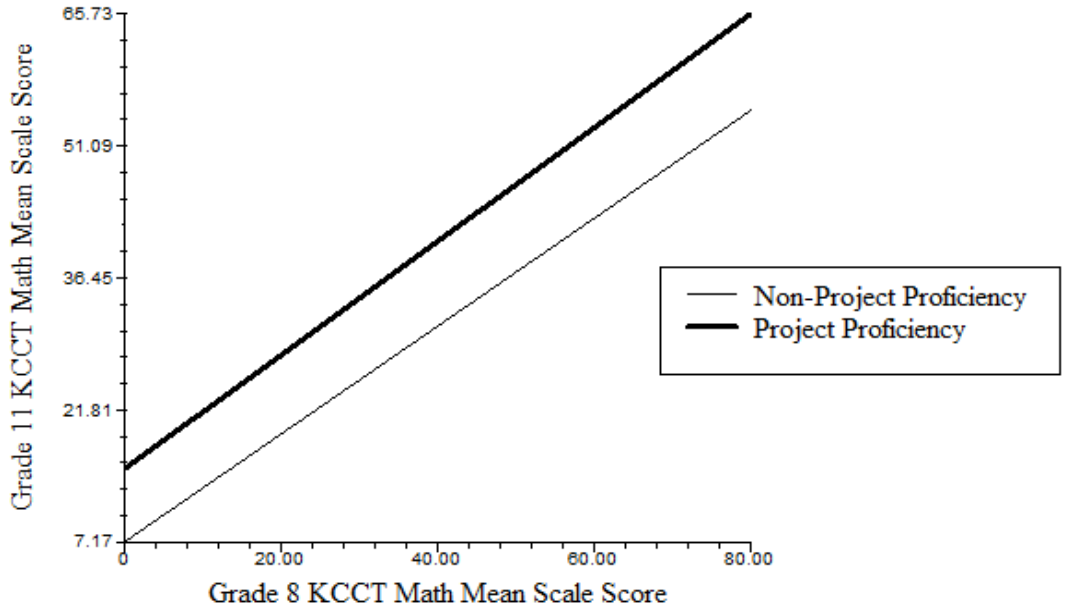
Note. *$p < .05$. **$p < .01$. ***$p < .001$. Model 1: Means-as-Outcomes. Model 2: Random-Coefficient. Model 3: Intercepts- and Slopes-as-Outcomes.

Random-coefficient regression model. The random-coefficient regression model was used to determine the average regression equations, examine variance in regression equations, and correlate intercepts and slopes. The Level 1 model contains the distribution of 11Math achievement through an intercept (β_{0j}) and slope (β_{1j}) in addition to average StudentSES (γ_{10}) and 8Math (γ_{20}) regression slopes across classrooms at Level 2. A highly significant association was found between 8Math ($\gamma_{20} = .60$, $t = 43.01$, $p < .001$) and 11Math, while a significant association was not found between StudentSES ($\gamma_{10} = -.54$, $t = -1.08$, $p = .28$) and 11Math (Table 11). Chi-squared tests identified highly significant differences between classrooms at the $p < .05$ level (Appendix 1). However, statistically significant differences were not found among StudentSES and 8Math slopes.

Similar to the means-as-outcomes model, the researchers calculated a proportion of variance explained for the random-coefficient model, (203.37 – 122.21)/203.37 = .40

(Table 12). StudentSES and 8Math reduced within-classroom variance by 40%. Since PP and ClassSES accounted for 32% of between-class variance in the means-as-outcomes model, it is clear that the association between prior achievement and 11Math performance was slightly stronger at the student level than at the classroom level. With the information obtained from this model, we can infer that a significant positive relationship between prior achievement and future performance exists; however, a statistically significant relationship between StudentSES and 11Math achievement was not present for students in this sample.

Figure 2. Grade 11 Math Achievement as a Function of Prior Achievement and Controlling for Classroom SES



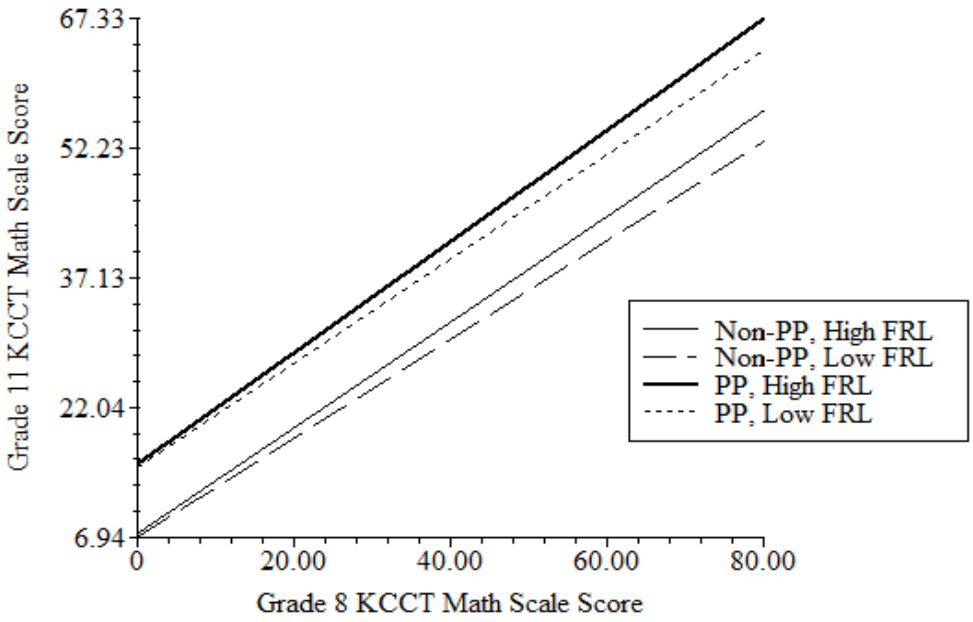
Intercepts- and slopes-as-outcomes model. The intercepts- and slopes-as-outcomes model in this study was designed to build a model to account for variability across classrooms and examine differences in the association between StudentSES, 8Math, ClassSES, and PP implementation. The intercepts- and slopes-as-outcomes

model combined the random-coefficient regression model at Level 1 and the means-as-outcomes model at Level 2 (Table 11). Though a positive correlation existed between ClassSES and mean 11Math achievement, the results were not statistically significant at the $p < .05$ level. The model suggested, however, that PP classrooms had significantly higher 11Math scores than non-PP classrooms when controlling for the effect of ClassSES, as $\gamma_{02} = 8.00$, $t = 6.04$, $p < .001$ (Appendix 1). These results are depicted graphically in Figure 2. No significant relationships existed between levels of ClassSES and PP on 11Math achievement on StudentSES slopes. PP classrooms produced slightly stronger 8Math slopes than non-PP classrooms; however, the results were not statistically significant. There was a tendency for classrooms with high percentages of low SES students to have significantly larger negative 8Math slopes as, $\gamma_{21} = -.16$, $t = -2.17$, $p < .05$ (Figure 3).

With the results obtained from the intercepts- and slopes-as-outcomes model, proportion of reduction of variance indices were calculated for the intercepts and slopes in the final model. For 11Math achievement (β_{0j}), where variance in the unconditional model had been 90.02, the residual variance was 20.39. The addition of ClassSES and PP to the random-coefficient model explained an additional 37% [(32.23 – 20.39)/32.23= .37] of variance (Table 12). This residual variance in the final model (20.39) represents a 55% reduction in between-classroom variance from the unconditional ANOVA model (90.02). With regard to 8Math slopes, the between-group variance present in the random coefficient model was .45 and the variance measure in the intercepts- and slopes-as-outcomes with the addition of ClassSES and PP to the model was .50. The proportion of variance explained in the later model -.11 [(.45 - .50)/.45 = -.11] which suggests that

these variables actually increased variance in slopes 11%. No reductions in variance were found in 8Math (β_2) slopes as a result of ClassSES and PP implementation, as variance component estimations in both the random-coefficient and intercepts- and slopes-as-outcomes models were zero.

Figure 3. Grade 11 Math Achievement as a Function of Prior Achievement, Project Proficiency and Classroom SES. FRL = Free or Reduced Lunch Eligible.



The deviance statistic for the intercepts- and slopes-as-outcomes model was 18924.94, representing reductions of 1415.98, 1362, and 120.19 over the unconditional, means-as-outcome, and random-coefficient regression models, respectively. In addition to overall variance reductions observed in the final model, these reductions in deviance represent improvements in model fit and verified that the intercepts- and slopes-as-outcomes model was a significant improvement over the means-as-outcomes and random-coefficient regression models (Kreft & De Leeuw, 1998).

To examine further the impact of PP on state mathematics achievement, a cross-tabulation of 8Math and 11Math achievement in the 2010 and 2011 groups revealed that 16.77% of students who failed to reach proficiency in grade 8 met the state-established benchmark in 2010-11, as compared to 4.64% in the 2009-2010 cohort (Table 13). In addition, the percentage of students who were at proficient or higher on 8Math and scored below proficient in 11Math decreased from 13.84% students in 2010 to 5.51% in 2011. These results suggest that the changes in teaching, assessment, and interventions in PP increased the number of students reaching AYP benchmarks and ameliorated drops in performance during high school.

Table 13

Students Meeting Grade 11 Proficiency Benchmarks Based on Prior Achievement and Project Proficiency Implementation

		Non-Proficient Grade 11		Proficient Grade 11	
	Cohort	n	%	n	%
Non-Proficient Grade 8	NPP	812	69.82	54	4.64
	PP	723	56.13	216	16.77
Proficient Grade 8	NPP	161	13.84	136	11.69
	PP	71	5.51	278	21.60

Note. NPP = Non-Project Proficiency (N = 1163). PP = Project Proficiency (N = 1288).

HLM Models- Social Studies as Outcome

In order to validate the use of 11SocStu as an outcome measure in the nonequivalent control group design, a Pearson correlation coefficient was calculated for the relationship between 11Math and 11SocStu for the 2010 and 2011 cohort groups. A strong positive relationship was found for the 2010 (r (1161) = .63, $p < .01$) and 2011 (r (1286) = .61, $p < .01$) cohorts, indicating a significant linear relationship between the two variables. In both cohorts, students who scored higher on 11Math tend to score high

on 11SocStu. Tables 14 and 15 present results of the four HLM models using 11SocStu as the outcome variables.

In the final HLM model, PP had a statistically significant impact on mean achievement and Student SES and 8SocStu slopes at the $p < .05$ level (Table 14). Classroom SES did not have a statistically significant impact on mean social studies achievement and student SES and prior achievement slopes in the final model, yet differences were found in the means-as-outcomes model at the $p < .001$ level. PP had a statistically significant negative influence on prior achievement slopes as $\gamma_{22} = -.09$, $t = -2.47$, $p < .05$ (Appendix 1). Chi-squared tests revealed statistically significant differences among classrooms in achievement in all four models at the $p < .05$ level; however, differences were not observed in student SES and prior achievement slopes (Appendix 1).

Table 14

Fixed Effect HLM Models: Grade 11 Social Studies as Outcome

Fixed Effect	Unconditional Model	Model 1	Model 2	Model 3
Classroom Means				
Intercept, γ_{00}	27.49***	26.64***	10.64***	8.43***
ClassSES, γ_{01}		-12.84***		-1.12
PP, γ_{02}		1.57		3.32*
StudentSES-achievement slopes				
Intercept, γ_{10}			-.90	-2.00*
ClassSES, γ_{11}				-.50
PP, γ_{12}				2.64*
8SocStu-achievement slopes				
Intercept, γ_{20}			.58***	.62***
ClassSES, γ_{21}				-.10
PP, γ_{22}				-.09*

Note. *$p < .05$. **$p < .01$. ***$p < .001$. Model 1: Means-as-Outcomes. Model 2: Random-Coefficient. Model 3: Intercepts- and Slopes-as-Outcomes.

In the unconditional social studies model, 25% of total variance was present between classrooms, as 58.10/(58.10 +170.97) = .25 (Table 15). The combination of PP and ClassSES reduced variation between social studies classrooms in the means-as-outcomes model by 11%, as (58.10 – 51.50)/58.10 = .11 and reduced total variance by 2% (.23). In the random-coefficient regression model, StudentSES and 8SocStu reduced within-classroom variance by 30%, as (170.97 – 116.40)/170.97 = .30 (Table 15).

Table 15

Random Effect HLM Models: Grade 11 Social Studies as Outcome

Random Effect	Unconditional Model	Model 1	Model 2	Model 3
Classroom Mean, u_{0j}	58.10***	51.50***	31.73*	31.17*
StudentSES-achievement slope, u_{1j}			8.65	6.90
8Math-achievement slope, u_{2j}			.01	.01
Level-1 effect, r_{ij}	170.97	171.37	116.40	116.35
Intraclass Correlation	.25	.23	.21	.21
Proportion Reduction in Variance Explained		.11	.30	.02

Note. *p < .05. *p < .01. ***p < .001. Model 1: Means-as-Outcomes. Model 2: Random-Coefficient. Model 3: Intercepts- and Slopes-as-Outcomes.

When we recall that PP and ClassSES accounted for 11% of between-class variance in 11SocStu in the means-as-outcomes model, it is clear that the association between student SES and prior achievement is nearly three times as strong at the student level than at the classroom level. In the final model that controlled for student SES and prior achievement at the student level and PP implementation and classroom SES at the classroom level, variance reduced by 2% over the random-coefficient regression model, with 21% of total variance present between classrooms. The deviance statistic for the intercepts- and slopes-as-outcomes model was 18815.15, representing reductions of

1057.83, 1036.03, and 41.38 over the unconditional, means-as-outcome, and random-coefficient regression models, respectively (Appendix 1).

Table 16

HLM Models: Percent of Variance Estimates for Mathematics and Social Studies

	Between-Classroom Variance (ICC/CICC)		Proportion Reduction in Variance Explained	
Model	Mathematics	Social Studies	Mathematics	Social Studies
Unconditional Model	31	25		
Model 1	23	23	32	11
Model 2	18	21	40	30
Model 3	14	21	37	2

Note. ICC = intraclass correlation. CICC = conditional intraclass correlation. Model 1: Means-as-Outcomes. Model 2: Random-Coefficient. Model 3: Intercepts- and Slopes-as-Outcomes.

HLM Model Comparison- Math versus Social Studies

With regard to fixed effects, mean classroom scores in the unconditional model were similar on both 11Math (28.50) and 11SocStu (27.49) measures. The unconditional models reported that more between-classroom variance existed in 11Math (90.02) than in 11SocStu (58.10), and student-level variance was higher in 11Math (203.37) than in 11SocStu (170.97). Table 16 presents a comparison of between-classroom variance and variance reduction in the HLM models using 11Math and 11SocStu as outcome variables. Between-classroom variance was slightly higher in 11Math in the unconditional model, with 31% between-classroom variance in 11Math and 25% between-classroom variance in 11SocStu. When comparing the results of the two regressions with means-as-outcomes models, PP and ClassSES had a considerable impact on predicting 11Math scores, as including the two variables reduced between-classroom variance in math (32%) as opposed to (11%) in social studies. The random-coefficient regression model comparison revealed that prior academic achievement had a strong effect on variance in

mean achievement within classrooms for both math (40%) and social studies (30%). StudentSES did not have a statistically significant effect on mean achievement in either model.

When PP and ClassSES variables were added to prior achievement and StudentSES to create the final model, within-class variance decreased in math classrooms by 37% and social studies variance by 2%. Prior achievement slope variability was not reduced by adding PP and ClassSES into either intercepts- and slopes-as-outcome model. Finally, after controlling for prior achievement and student SES at Level 1 and participation in Project Proficiency and class SES at Level 2, 14% of variance existed between mathematics classrooms and 21% variance existed between social studies classrooms, representing a 55% variance reduction in mathematics from the original variance estimate of 31% and a 16% reduction in social studies from the original variance estimate of 25% in the unconditional models.

Discussion

Using two years of demographic and test score data from students in one school district, this study used two-level HLM models to analyze the effects of PP on mathematics achievement. Under the PP initiative, teachers collaborated in using common formative and summative assessments to guide students to competency on key standards. Due to the hierarchical nature of education data and the inability of OLS regression models to analyze student and classroom levels at the same time, HLM was employed to evaluate the effect of PP on student achievement independently of the effects of other variables operating at either the student or classroom level. HLM analysis enabled the researchers to analyze student-level, classroom-level, and cross-level

relationships (Raudenbush & Bryk, 2002). The HLM analysis controlled for prior academic achievement and individual SES status at the student level and cumulative SES and PP participation at the classroom level. HLM models were used to create an unconditional model that examined mean student achievement with no controls, a model that controlled for student-level factors, a model that controlled for classroom-level variables, and a final conditional model that controlled for student-level and classroom-level variables simultaneously.

With regard to the first research question, the HLM analysis revealed that PP had a significant impact on mathematics achievement. PP accounted for a statistically significant increase of 8 mean scale score points in 11Math in a model that controlled for classroom SES and prior student achievement. This equals nearly one-half of a performance level on the state assessment (Kentucky Department of Education, 2008). The study revealed that prior student achievement played a major role in student outcomes as 40% of student-level variance in 11Math was accounted for by adding prior student achievement to the HLM model. Despite the strong correlation between prior achievement and student outcomes, PP appeared to have increased the number of students reaching proficiency benchmarks.

With regard to the second research question, results from the HLM models were consistent with the wide body of research stating that school socioeconomic status has a greater influence on student achievement than an individual student's SES. The study found that while prior mathematics achievement was a powerful predictor of performance, no significant relationship was found between individual student SES and mathematics achievement, and classroom SES had a significant influence on student

achievement. We found evidence in the HLM analysis that classroom practices accounted for 31% of variability of mathematics scores between classrooms in our unconditional model. This was consistent with similar HLM analyses conducted by Borman and Dowling (2010); Bryk, Sebring, Allensworth, Luppescu, and Easton (2010); and Hattie (2003). PP implementation further reduced differences in mathematics mean scale scores as between-classroom variance decreased by 25% in the means-as-outcomes model and 55% in the final model. These results dwarf between-class variance decreases of 11% and 16% in social studies, respectively. In the 110 classrooms studied, significant reductions in between-classroom variation in mathematics achievement suggested that PP impacted instruction at the classroom level, as only 14% of differences in mean scale scores existed between-classrooms in the final model. When examining the effects of PP in a district-wide effort to improve test scores and move reform to scale, reductions in between-classroom variance such as this imply that PP positively changed classroom instructional practices.

Conclusion

Returning to the questions posed at the beginning of this study, it is now possible to state that PP positively increased student achievement and reduced the variance of student performance and classroom compositional effects on state mathematics assessments. The present study confirmed previous findings and contributes additional evidence that quality teaching does have an impact on student achievement, despite the influence that classroom SES and prior achievement have on student learning. In the current high-stakes education environment that requires dramatic increases in student achievement for all students, regardless of race, SES, or disability, PP demonstrated the

potential to provide students with the ability to reach their academic goals in route to becoming productive members of society. PP gave an urban school district a scalable reform design capable of meeting the demand for dramatic increases in student performance. The results of the study suggest that PP improved student performance in mathematics and decreased variability in achievement across classrooms. The significant reduction in between-classroom variance found in this study suggests that PP was widely adopted by teachers and produced fundamental changes in classroom practices necessary to reach scale (Coburn, 2003).

The current study demonstrated that mathematics achievement increased as a result of PP implementation. What is less clear is the fidelity of PP implementation in the classrooms studied. Quantifying implementation levels provides researchers with an additional control variable in research studies (Muñoz, Guskey, & Aberli, 2009; O'Donnell, 2008). Though research by Stringfield et al. (2008) tells us that reform efforts are most successful when schools and teachers have efficacious interactions with the reform design, the current study lacked measures or analysis of the quality of classroom instruction or fidelity to the PP design. The researchers relied on school administrators to identify PP classrooms for the study and did not ask principals or teachers to quantify PP implementation levels (i.e. Datnow, Borman, Stringfield, Rachuba, & Castellano, 2003; Supovitz and Weinbaum, 2008).

A second limitation of the study centers on changes in teaching. Specifically, did PP cause changes in teaching or did the fact that 10 of 21 schools in the district were designated as PLA create a "fire in the belly" of teachers to change practices and demand more from themselves and their students? With eight of the 11 schools in the study

labeled as PLA, one of the study's threats to external validity is related to the relationship between the population, the sample, and the study's generalizability. A selection-history threat (Trochim & Donnelly, 2008) is present in that schools in the 2010 cohort were not directly experiencing the accountability, scrutiny, and consequences of being labeled PLA. The threat of sanction was present, but schools were not undergoing intended-to-be-transformational activities. The current study operated on the premise that PP was the reform effort that spurred changes in teaching, assessing, and intervening with students experiencing difficulty. Given this context, critics may question the ability of the study to be generalized to schools not under NCLB sanctions or that do not contain student populations resembling those in the study.

Implications for Policy, Practice, and Future Research

The evidence from this study suggests that PP positively impacted student achievement and improved classroom teaching. In general, therefore, it seems that the elements of PP have the potential for success in content areas other than mathematics. A first implication is that school and district leaders looking to improve academic performance should considers reforms like PP that clearly identify learning targets, utilize formative assessments, and provide mechanisms that guarantee student learning. In this context, many more students demonstrate competency, receive supports, and are not left behind when they fail to understand academic concepts. PP created conditions that were intended to shift ownership of learning from solely on students to a combination of the student, the teacher, and the teacher's PLC. With NCLB, RTTT, and SIG requirements measuring school effectiveness through proficiency measures on state assessments, proactive and systemic approaches such as PP should be implemented to improve student

achievement. Reforms like PP have the potential to dramatically increase the number of students reaching proficient levels and create the conditions for scale up and sustainability identified by Coburn (2003) and Stringfield et al. (2008).

In his study of professional development and changes in teacher practices, Guskey (2002) observed that teacher beliefs and attitudes adjust only after observing positive student outcomes. Guskey (2002) asserted, "The key element in significant change in teachers' attitudes and beliefs is clear evidence of improvement in the learning outcomes of their students" (p. 384). The strong connection between effective teachers and student achievement creates the need for instructional innovation, like PP, to make changes in teaching that lead to improvement in student outcomes and produce conditions for the program to move to scale.

PP set out to improve student achievement through fundamental changes in teaching practices and assessment and intervention systems. Teacher efficacy strengthened through district-supported professional development and a strong emphasis placed on professional collaboration. The combination of structures, systems, and supports reflect the research on teacher quality, collaboration, instructional practices, and professional development researched by Aaronson et al. (2007), Guskey (2002), Hattie (2003), Odden et al. (2004), and Stronge (2010) that yielded improved instructional practices and student performance.

Conditions for scale up and sustainability are strengthened through strong alliances between a school district and their schools (Bryk et al., 2010; Datnow, 2005; Fullan, 2000; Fullan, Bertani, & Quinn, 2004; McLaughlin, 1990; Oxley & Luers, 2010). A second implication validates the importance of and need for strong district-school

relations, as these bonds are usually given short shrift (Rorrer, Skrla, & Schuerich, 2008; Smith & O'Day, 1991). District leaders in Jefferson County provided unprecedented support to PP schools by creating curriculum materials, modifying data systems that track individual student progress towards key standards, and providing professional development in formative assessment (Stiggins et al., 2004) and professional collaboration (DuFour et al., 2004). Commenting on the role of districts in scale up efforts, Sanders (2012) observed that district supports that include targeted professional development and structures that engage schools in dialogue regarding professional practice can "significantly influence the quality with which external reforms are implemented and sustained over time" (p. 157). In their narrative synthesis of the role of school districts in educational reform, Rorrer et al. (2008) challenged the conceptualization of school districts and proposed that school districts are an "organized collective constituted by the superintendent; the board; the central office-level administration; and principals, who collectively serve as critical links between the district and the school for developing and implementing solutions to identified problems" (p 311). Through strong school-district relations, PP provided the structures and supports necessary for scale up, and school districts that prioritize and foster such relationships have the potential to move reform to scale.

Future research on PP should include examining implementation fidelity through classroom observations, data review, and stakeholder interviews. Subsequent HLM analysis containing controls for program implementation fidelity would give researchers the ability to more deeply examine the four distinct elements of PP, determine which ones have the most impact on student achievement, and what relationship these elements

have on one another. The proposed mixed methods research (Creswell & Plano-Clark, 2003; Tashakkori & Teddlie, 2003) would provide researchers and policymakers a more critical analysis of PP implementation. Future research would also shed a light on the following barriers to reform identified by Berends, Bodilly, and Kirby (2002), Datnow (2005), Datnow and Stringfield (2000), Slavin et al. (1996), and Stringfield and Datnow (1998): failure to implement reforms school-wide, lack of foresightedness on how reform models would fit school goals and changes in school, and counterproductive district policy as barriers to reform. In these complex and sophisticated contexts, further examination of the school-district relationship would provide insight into program implementation, allow for analysis of challenges to reform, and recommend actions to strengthen them. Researchers, policymakers, and stakeholders can use this information to further determine if PP yields the academic gains necessary for to take large numbers of students to proficiency and contains the dimensions of scale identified by Coburn (2003) to sustain instructional innovation.

THE STUDENTS IN FRONT OF US: REFORM FOR THE CURRENT GENERATION OF URBAN HIGH SCHOOL STUDENTS

Navigating unprecedented sanctions, high school educators in the Jefferson County Public Schools (JCPS), a large urban district in Louisville, Kentucky, had one year to move dramatic numbers of students' scores on statewide tests. The charge was to raise the scores to levels of proficient performance in reading and mathematics and establish sustainable reform. The mandate resulted from the pursuit of Race to the Top funds by Kentucky legislators and policymakers. The Kentucky state legislature aligned state statutes with federal turnaround models for schools identified as persistently low-achieving (PLA), generating additional layers of school accountability for schools scoring in the bottom five or 5% of the state in reading and math (Persistently Low Achieving School, 2010). With JCPS test scores that ranked as the lowest in the state, state education officials identified ten of 21 JCPS high schools for sanctions and state monitoring and placed two additional schools on probation. The sanctioned schools collectively averaged 59% minority and 80% qualifying for free or reduced-price lunch (FRL). All faced possible removal of their principals and at least 50% of their faculties. In the fall of 2010, amid the landscape of uncertainty and low morale, exacerbated by challenges of moving forward large numbers of disadvantaged students found in large urban districts, JCPS developed and implemented a plan to rapidly and uniformly impact the academic performance of high school students in their PLA schools.

In September 2011, results of the Kentucky Core Content Assessment (KCCT) yielded the largest high school reading and math gains in JCPS history, prompting celebration, while also challenging JCPS educators to replicate and sustain the progress. All 21 JCPS high schools gained in reading and math proficiency. The JCPS PLA school averaged a 14% gain compared with a 5% state gain in reading and a 17% gain compared with a 6% state gain in math. However, even with widespread gains, five high schools remained in or near the bottom five or 5% of the lowest performing schools in the state. To escape PLA status, JCPS educators needed to maintain the momentum of its high school reform and strategically address its most critical student learning gaps. Improving overall averages in student proficiency can mask the need to move each individual student significantly forward. Particularly in PLA schools, genuine and sustainable reform efforts must help transform large percentages of low achieving students – generally low-income, minority, and transient students – to proficient performers (Zavadsky, 2009).

The purpose of this study was to examine the impact of a successful JCPS high school reform model on the academic achievement of its most at-risk students. To achieve this purpose, the researchers used a comparative study to answer the question:

> When compared with similar students at-risk of dropping out, do statistically significant differences exist in gains in KCCT math scale scores between the 8th and 11th grades for students participating in the JCPS high school reform model?

High-Stakes Accountability

Concerned that declining educational standards were producing American mediocrity (United States Department of Education, 1983) and an ill-equipped 21st century workforce, Americans have demanded high school reform. To counter this

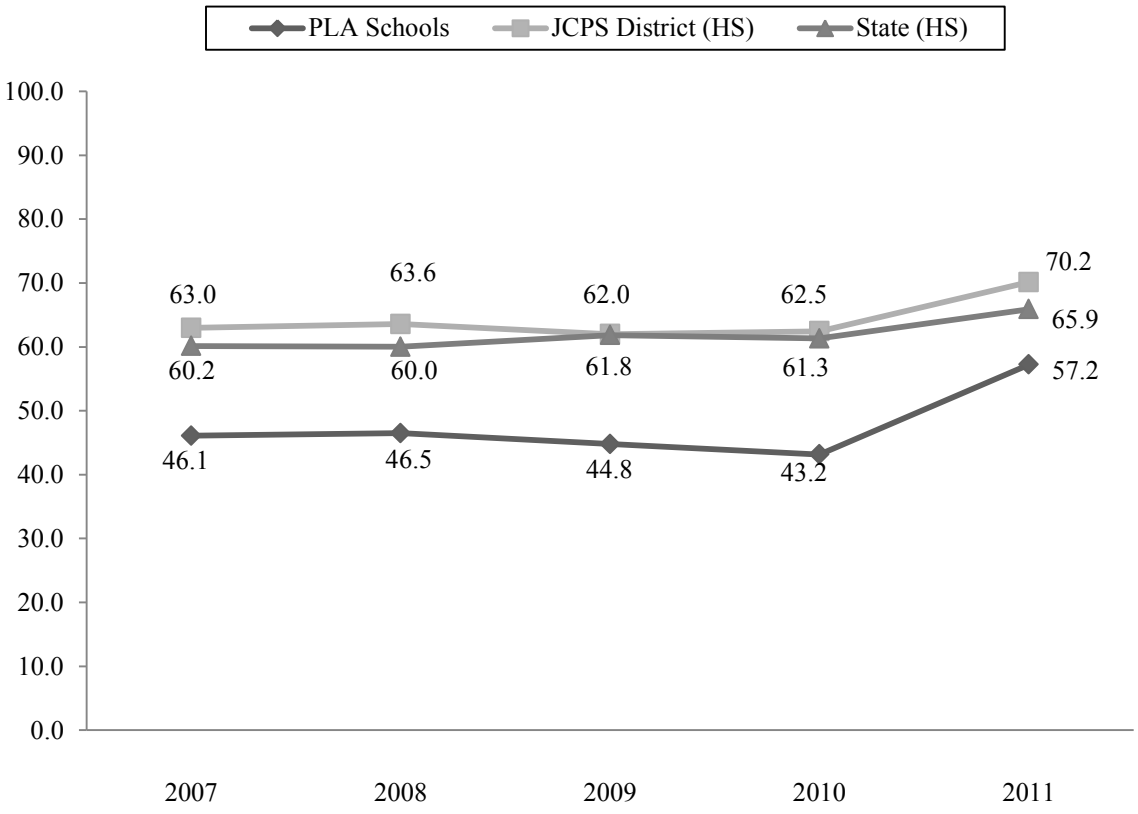
perceived lack of adequately rising educational standards, policymakers established high-stakes accountability measures for public schools (School Improvement Fund, 2010) and business and civic leaders encouraged school districts to emulate international education practices (National Center on Education and the Economy, 2008). Despite a variety of school reforms, Lasky et al. (2005) affirmed "the nation's best measure of school-age academic achievement, the National Assessment of Educational Progress (NAEP), has, over its three decades of existence, documented very little progress toward demonstrable improvement in student performance" (p. 28). Even though recent high school results demonstrated an increase in proficiency in reading and math from 2005 to 2009, reading scores fell below 1992 results and achievement gaps among subgroups of students remained unchanged (NAEP, 2011). Unfortunately, although graduation rates have trended slightly upward this past decade (Bruce, Bridgeland, Fox, & Balfanz, 2011), little evidence exists that indicates school districts have moved high school reform to scale and obtained significant gains in student achievement (Balfanz & Legters, 2004a; Bryan, Klein, & Elias, 2007; Earl, Torrance, & Sutherland, 2006; Stringfield & Datnow, 1998).

JCPS High School Reform: Project Proficiency

In 2010, after the Kentucky Department of Education (KDE) identified 10 PLA high schools in JCPS and several others on the verge of PLA status, JCPS district leaders developed a high school reform strategy, Project Proficiency (PP) (Jefferson County Public Schools, 2011a), to accomplish three challenging goals: generate substantial gains in reading and math proficiency, achieve results in a short amount of time, and propagate the reform throughout all PLA high schools. First, JCPS needed a plan to dramatically increase proficiency test scores in reading and math since previously established JCPS

high school reforms of leadership development, school choice, smaller learning communities, inquiry-based curriculum, data-tracking systems, and equitable funding had yielded positive, but only incremental gains. From 2007 to 2010, many of the JCPS PLA and near-PLA high schools had sporadic increases in average proficiency in reading and in math, but their scores still ranked in the state's bottom five or 5% (Figures 4 and 5). JCPS leaders needed a systemic plan for radical change in instructional practice to achieve dramatic gains across all PLA high schools.

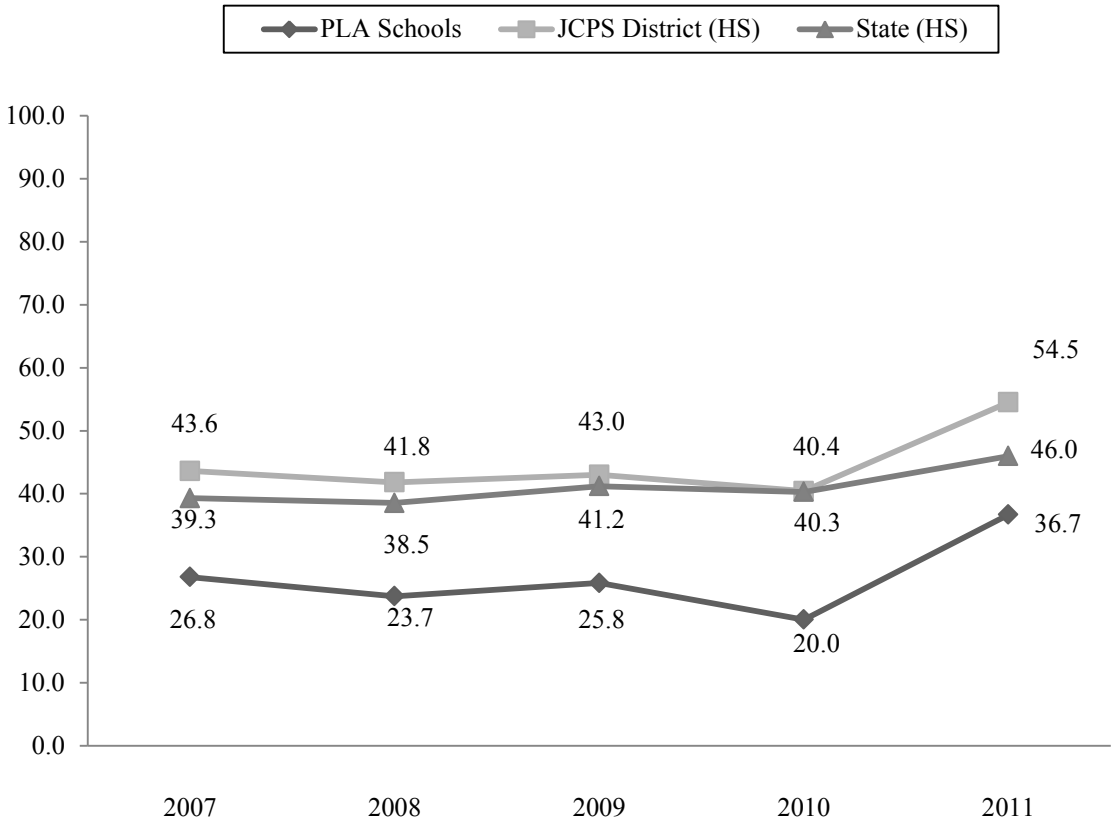
Figure 4. KCCT Reading Proficiency Trend Data 2007-2011



Second, JCPS leaders required a plan for quick improvement of PLA high schools. Each year, KCCT results potentially brought unprecedented sanctions of staff removal, charter take-over, and school closing for the "lowest-achieving schools in

improvement, corrective action, or restructuring in a state" (School Improvement Fund, 2010). Amid the PLA landscape of low teacher morale, diminished school credibility, and varying student performance levels of its large numbers of disadvantaged students, JCPS leaders sought a system to prepare students swiftly to perform proficiently on an imminent KCCT assessment.

Figure 5. KCCT Math Proficiency Trend Data 2007-2011.



Third, the volume of PLA schools forced JCPS central office leaders to move PP to scale through an intentional shift of ownership of PP principles to multiple settings (Coburn, 2003; Stringfield & Datnow, 1998). High percentages of PLA schools and at-risk high school students, typical of large urban districts (Balfanz et al., 2007; Darling-

Hammond et al., 2005; Earl et al., 2006; Stringfield & Yakimowski-Srebnick, 2005), compelled JCPS leaders to move beyond targeted reform for students with near proficient scores at selected schools to broad reform for every high school English and math classroom in ten PLA schools.

Combining effective practices gleaned from previous JCPS high school initiatives and school turnaround strategies, JCPS administrators and teacher leaders developed and launched PP across PLA high schools for the 2010-11 school year. The guiding principle of ensuring learning, or guaranteeing student competency in key standards, drove systemic reform toward the three JCPS reform goals of dramatically increasing student proficiency, accomplishing turnaround in one school year, and fostering propagation of PP to scale. First, setting as a goal the guaranteeing of competency for *all* students leveraged radical instructional changes in standards-based teaching and assessment of student work. Second, ensuring student understanding of fewer standards on a fixed grading-period schedule accelerated learning of key content regardless of a student's starting point or background knowledge. Third, PP moved to scale within and across schools as teachers, operating within professional learning communities (PLCs) (DuFour, 1998), collectively owned student results and benefited from one another's expertise to "co-construct" (Datnow & Stringfield, 2000, p. 188) lessons, tasks, and interventions to guarantee student competency of each key standard.

JCPS established conditions through the practices of PP for balancing instructor commitment to teaching with their ownership of student learning. The system provided a practical framework for teachers to guide 100% of their students to demonstrate competency in three key standard categories in English 2, Algebra 1, Geometry, and

Algebra 2 each grading period before taking a district end-of-grading-period proficiency assessment. Competency was not proficiency or mastery, but a level of performance that teachers viewed as qualifying a student for the proficiency assessment.

Guided by the goal of guaranteed competency for each student, each school navigated implementation in a unique way to put the 4-part PP framework in place: (1) focus on three key standard categories each grading period, (2) create tasks through which students could demonstrate competency of key standards, (3) collectively ensure or guarantee student competency in all three key standards by the end of the grading period, and (4) implement fail-safe assessments, requiring recovery for students scoring below 80%. Teams of teachers who taught the same course formed PLCs that collaborated weekly to discuss data on student progress toward competency and examined student work samples for instructional implications. Through re-teaching, differentiated instruction (Tomlinson & McTighe, 2006), redesign of student tasks to assess student competency (Stiggins, 2008), responsive interventions (Fuchs & Fuchs, 2006), and standards-based grading (Guskey, 2009; Lekholm & Cliffordson, 2008; Marzano, 2010), teachers collectively sought to ensure student competency for each key standard and acquired a shared knowledge base of effective instructional practices for subsequent grading periods (Abbott & Fisher, 2011; Allen & Blythe, 2004; DuFour, DuFour, Eaker, & Karhanek, 2004).

The 2011 KCCT scores reinvigorated JCPS staff, inspired many students, and provided the community with the hope that effective reform had turned around the PLA schools. However, JCPS district officials questioned whether PP could generate continuous improvement beyond one year of implementation. Stringfield and Datnow

(1998) asserted that reform efforts aimed at urban districts have lacked systemic sustainability. Payne (2008) added that large urban districts were primarily challenged with systemically moving significant numbers of at-risk students forward and concluded that while urban districts tinkered with a variety of researched-based effective strategies for improving performance of disadvantaged students, "we don't know how to implement these things with fidelity at scale" (p. 94). Significant increases in proficiency by JCPS at-risk students could provide credible evidence that PP overcame the odds to create outcome-raising high school reform at scale.

Targeting Reforms Toward Potential Dropouts

If future American prosperity depends upon elevating each student's preparedness for a rapidly changing and globally competitive environment, then urban districts' reforms must impact their high schools' large numbers of students who drop out (America's Promise, 2010). Dropouts tend "to be unemployed, living in poverty, receiving public assistance, in prison, on death row, unhealthy, divorced and ultimately single parents with children who drop out of high school themselves" (Bridgeland, Dilulio, & Morrison, 2006, p. 2). With low-skill work increasingly outsourced internationally and commanding low wages intra-nationally, dropouts have affected the general economy by limiting their wage-earning power, resulting in a loss of billions of dollars in family incomes and American tax revenue (Aud et al., 2010; European Commission, 2009; Gordon, 2009; Harlow, 2003; Land & Legters, 2002; Levin, Belfield, Muennig, & Rouse, 2007; Wald & Losen, 2003). America's large numbers of dropouts have highlighted the problem. Balfanz and Legters (2004b) reported, "between 1993 and 2002, the number of high schools with the lowest levels of success in promoting

freshmen to senior status on time, a strong correlate of high dropout and low graduation rates, increased by 75%" (p. 4). To preserve and strengthen America's competitiveness, urban high school reform must convert potential dropouts into well-educated workers in a knowledge-based economy (Toch, 2003).

Middle school prevention: Too late for current high school students. Recent studies recommended that urban districts attack the dropout epidemic through early identification and prevention. For instance, in a study of on-time graduation characteristics of freshmen in the Chicago Public Schools, Allensworth and Easton (2005) found students who entered high school from the bottom quarter of their 8th grade were more than 40% off track to graduate by the end of freshman year. Balfanz, Herzog, and MacIver (2007) examined longitudinal data of 12,972 Philadelphia public school students and concluded that urban districts could use 6th grade individual factors of low attendance, failure of math or English, or suspensions to identify and prevent 60% of potential high school dropouts (Table 17). Additional studies of large urban districts have affirmed that students entering the ninth grade over-aged, chronically absent, from low-income families, who present poor behavior and low achievement scores have a higher dropout rate in high school and require middle school interventions (Alexander, Entwisle, & Kabbini, 2001; DeWit, Karioja, & Rye, 2010; MacIver, 2010; MacIver, Durham, Plank, Farley-Ripple, & Balfanz, 2007; Neild & Balfanz, 2006; Silver, Saunders, & Zarate, 2008; Zvoch, 2006). In fact, MacIver (2011), questioning the impact of high school adult advocacy on student engagement, attendance, on-time promotion, and graduation, found no significant effect on dropouts and concluded "relatively well-implemented strategies that are research-based to prevent dropout will

not necessarily yield positive effects unless systematically linked to a complete framework that begins at least in the middle schools" (p. 181). These studies suggest successful transition from middle school to high school as the key to significant reduction in high school dropout rates.

Table 17

Students' End-of-6th Grade Measures Predictive of 60% Dropout Probability

Below 80% Attendance
End-of-course failure in math
End-of-course failure in English
Suspended or low end-of-course conduct grade

Note. Students meeting at least one of these four criteria at the end of 6th grade have a 60% chance of dropping out of school. Adapted from "Preventing Student Disengagement and Keeping Students on the Graduation Path in Urban Middle-grades Schools: Early Identification and Effective Interventions," by R. Balfanz, L. Herzog, and D. MacIver, 2007, *Educational Psychologist, 42*(4), p. 227.

Unfortunately, middle school identification and prevention strategies for at-risk students provide no remedy for students already in PLA high schools or in districts not implementing early warning systems. In large urban districts, lagging indicators have arrived too late for high school educators to determine the effectiveness of previously implemented strategies and target struggling students for specific remediation (Mishook, Foley, Thompson, & Kubiak, 2008). This lack of prevention forces educators in large urban districts to own, transform, and prepare the students who are in the desks in front of them for proficient performance on annual state assessments (School Improvement Fund, 2010).

Whole school reform: Too slow and sporadic for the current generation of students. Lacking the luxury of middle school interventions to improve at-risk student performance, many urban district leaders have struggled to implement and sustain reform at scale. Organizational bureaucracy has compounded the problem, stifling accelerated

change required to reform the high percentage of low-performing urban high schools (Bryan et al., 2007). Despite pockets of possibility providing hope for scale up, Earl et al. (2006) argued, "there are no examples anywhere of successful whole district high school reform. There are a few high schools, here and there, that have improved significantly, but none as a group" (p. 126).

Neild, Balfanz, and Herzog (2007) recommended that urban high schools should develop a comprehensive set of strategies to monitor freshman progress toward graduation, coordinate staff collaboration and decision-making, and create a climate of success for disengaged students. However, urban districts have encountered issues with implementing each of these strategies. Providing supports and interventions for freshmen off track for graduation have yielded unreliable results based on a school's level of implementation. After examining the impact of Chicago Public Schools' reforms, Bryk, Sebring, Allensworth, Luppescu, and Easton (2010) emphasized the incredible difficulty of expanding a school's successful program for disadvantaged students into an organization-wide reform.

The Baltimore City Public School System (BCPSS) has replicated some of its successful reform elements such as freshman academy and targeted support of at-risk students, yet has not expanded its Talent Development Model to scale across the district despite significant increases in performance found at Patterson High School (McPartland, Balfanz, Jordan, & Legters, 1998). Perhaps leaders of the local education authorities in Chicago and Baltimore lacked a "set of knowable, replicable technologies for scaling up the promising programs and practices so that they can be used by relatively typical professional educators" (Stringfield & Datnow, 1998, p. 270).

Educational leaders concerned by the performance of public schools have attempted a myriad of initiatives to increase student achievement. Charter schools (Bilko & Ladd, 2006; Lubienski, 2003; Nathan, 1999) trimester schedules (Lybbert, 1998; Winn, Menlove, & Zsiray, 1997), career academies (Brand, 2009; Kemple & Willner, 2008; McLaughlin, 1990; Quint, 2006), and school choice (Hoxby, 2003; Levin, 2001) have been attempted, but none has proven dramatically successful.

Whole-school coordination of high school staff and supports that have generated successful outcomes have typically required time, organization, and resources that were valuable for long-term reform but impractical for producing short-term, district-wide increases in student achievement. In a longitudinal, mixed-method case study, Stringfield and Yakimowski-Srebnick (2005) found that in six years, the effects of accountability-driven reforms in BCPSS of testing, governance, and federal No Child Left Behind (NCLB) legislation generated a 13.1% gain in the high school graduation rate. In a study of the San Diego Unified School district, Darling-Hammond et al. (2005) found over five years that due to systemic improvements in principal instructional leadership, higher level course offerings to all high school students, and extended learning opportunities, the percentage of high schools that met state and subgroup NCLB targets increased from 19% to 56%. The Chicago Public Schools' reforms targeted students on track for graduation and decreased the dropout rate by more than 4% over eight years (Allensworth & Easton, 2005, 2007; Bryk, Sebring, Kerbow, Rollo, & Easton, 1998; Hess, 2003).

Almeida, Steinberg, Santos, and Le (2010) reported that the New York City Public schools established 42 high schools that graduated above average numbers of

"over-age and undercredited" students. In addition, the North Carolina New Schools Project launched by the governor and the North Carolina Education Cabinet in 2003 created schools that reported an average 1.3% smaller dropout rate than comparison schools across the state. Useem, Offenberg, and Farley (2007) studied the School District of Philadelphia's attempt to affect student outcomes through improving teacher certification and quality. Despite multiple improvements in the hiring process, they concluded the school placement process continued to make it difficult to "move fast in hiring the best and brightest in a timely way" (p. 20). Unfortunately, if at all, these programs and projects improved only a percentage of the schools over several years, and to recover from or avoid PLA status, urban districts must take high school reform to scale expeditiously with no time for a "pause button" (Bryan et al., 2007). Although some high school reforms have shown promise, Balfanz, Legters, West, and Weber (2007) advised that even leading-edge reforms could take four years to move struggling students in PLA schools to proficient performance.

Improving student perceptions: Too little for dramatic gains. Interventions specifically designed to improve desired academic skills might improve academic self-concept, or the belief in one's academic ability. In a study of 1,211 secondary students in Australia, researchers (Bodkin-Andrews, O'Rourke, Dillon, Craven, and Yeung, 2009) found that student levels of academic self-concept predicted measures of school disengagement. Marsh and Craven (2006) found that self-concept of academic abilities yielded stronger student outcomes than self-esteem. In fact, based on a longitudinal study of five waves of survey, grade-point average, and eventual educational attainment data of over 2000 tenth graders, Marsh and O'Mara (2008) developed a "reciprocal

effects model" (p. 549) and established that increased academic self-concept led to improved performance, which cyclically led to further increases in academic self-concept. Despite evidence that students' perceptions of improved climate and a sense of belonging influenced their effort and performance and should bolster high school reform (DeWit et al., 2010; Rumberger & Lim, 2008; Wilkins, 2008), student perception of peer, teacher, and emotional support typically decreases at the secondary level (Barber & Olsen, 2004; Furrer & Skinner, 2003; Lynch & Cicchetti, 1997; Marks, 2000; Reschly, Huebner, Appleton, & Antaramian, 2008; Whitlock, 2004). Not to diminish the goal to improve students' confidence in themselves within a caring environment, without simultaneous improvements in teaching, it alone is unlikely to generate "substantial improvements in student learning" (Bryk et al., 2010, p. 17).

The Students in Front of Us

With JCPS high school teachers vacillating between hope and despair as a result of intensified NCLB and state sanctions, JCPS leaders launched PP across all PLA schools and provided unprecedented supervision and support for educators to move dramatic numbers of students to proficiency in a matter of months. District administrative leaders, curriculum specialists, and resource teachers regularly observed and participated in teacher PLCs to ensure a focus on prescribed curriculum, data-driven decision-making about common assessment data, and responsive interventions for struggling students. However, district leadership balanced increased supervision with improved and more prompt support, providing technology, curricular resources, scheduling adjustments, ongoing training opportunities, and funding to address concerns that surfaced during PLC meetings.

National reform recommendations for our most disadvantaged students include middle school intervention (Balfanz et al., 2007), whole high school reform (Nunnery, 1998), and student-perception adjustments (Baker, 2006). Although necessary for the at-risk students sitting in the desks in front of us, months away from the next high-stakes assessment, none of these effective reforms offers a sufficient solution to help them. Future middle school prevention programs are, by definition, too late for high school students. Similarly, whole school reform efforts take years to achieve strong implementations, and would be too late for current, low-achieving high school students. Adjusting student attitudes alone delivers too little impact on reading and math test scores.

Previous JCPS reforms had produced modest incremental increases in student results, but post-PP KCCT results yielded unprecedented proficiency gains in reading and math across all PLA high schools. Sustainability of PP and each PLA school's eventual escape from sanctions depend on the specific impact of PP on at-risk students predicted to "fall short of the graduation path absent intervention" (Balfanz et al., 2007, p. 226). Examining the impact of PP on the academic performance of at-risk students may provide insight into effective, district-wide, urban high school reform for the students sitting in the desks in front of us any given day and year.

Method

Participants

Jefferson County Public Schools. JCPS enrolls approximately 100,000 students in 90 elementary schools, 25 middle schools, 21 high schools, and 20 alternative settings. The alternative settings include schools for pregnant teens, zero-tolerance offenders,

adjudicated students, dropout candidates, and state agency students. JCPS serves 14% of Kentucky's total student population and nearly 50% of its African-American students.

JCPS Students. The composition of the district's student body includes 56.5% white, 36% black, and 7.5% other students. More than half of JCPS students reside in single-parent homes and approximately 63% qualify for FRL.

The initial sampling frame for this study included all of the JCPS students who attended the district's 11 PLA or near-PLA high schools during the 2009-10 and 2010-11 school years. The sample was divided into two cohorts. The comparison group, the non-PP cohort, took the 11th grade math and social studies KCCT tests in 2010 without participating in PP. The treatment group, the PP cohort, took 11th grade math and social studies KCCT tests in 2011 after participating in PP.

We narrowed each identified cohort sample to include only students with corresponding 8th and 11th grade math and social studies KCCT test scores. The non-PP cohort students had corresponding scores from 2007 and 2010 and the PP cohort students had corresponding scores from 2008 and 2011. Through purposive, nonprobability sampling (Trochim & Donnelly, 2008), we further reduced both cohorts to students who finished their 6th grade year with least one of four dropout-predictive criteria researched by Balfanz et al. (2007). The final non-PP cohort included 241 2005 6th grade students, and the final PP cohort included 264 2006 6th grade students who met the same criteria (Table 18).

Table 18

Cohort Characteristics of Students At-Risk of Dropping Out

	Non-PP Cohort		PP Cohort	
Characteristic	n	%	n	%
Race/ethnicity				
White	64	26.56	79	29.92
Black	170	70.54	173	65.53
Other	7	3.90	12	4.55
Gender				
Male	166	68.88	172	65.15
Female	75	31.12	92	34.85
Free/reduced lunch	189	78.42	212	80.30
ECE	58	24.07	67	25.38
ESL	5	2.07	13	4.92

Note. ECE = Exceptional Child Education (special education eligible). ESL = English as Second Language.

Measures

We used math and social studies KCCT test scores as our measures of student achievement. We used mean scale scores to compare corresponding KCCT 8th grade and 11th results. However, initial 8th grade means ranged from three to six points higher than corresponding 11th grade means (Table 19).

Table 19

KCCT Mean Scale Scores for 8th and 11th Grade Math and Social Studies

Cohort	Math	Social Studies
Non-PP Cohort		
2007 8th Grade	39	41
2010 11th Grade	36	35
PP Cohort		
2008 8th Grade	41	42
2011 11th Grade	37	36

To ameliorate differences between KCCT 8th grade and 11th grade mean scale scores, we mean-centered the scale scores for each student around the appropriate annual state mean scale score. We determined mean-centered scale scores by subtracting the

state mean scale score from each student's scale score. To promote test security, Kentucky administers multiple versions of the KCCT. The KCCT Test Technical Guide (Measured Progress, 2009) identifies Cronbach's alpha measures to report internal consistency. The eighth grade mathematics and social studies tests each consisted of six (6) test versions (α = .89), and the 11th grade mathematics and social studies tests each consisted of 6 test versions (α = .90). Item and description indices were identified for each test version and converted to mean scale scores from 0-80. For state and NCLB reporting purposes, Kentucky mean scale scores are divided into four performance level descriptors: novice, apprentice, proficient, and distinguished (Table 20).

Table 20

KCCT Mean Scale Score Range and Performance Descriptors

Math and Social Studies KCCT Test	Performance Level Description Range			
	Novice	Apprentice	Proficient	Distinguished
8th Grade	0-19	20-39	40-62	63-80
11thGrade	0-19	20-39	40-63	64-80

Design and Procedures

This study tested the assumption that PP was positively associated with KCCT math performance for students at-risk of dropping out of high school (Balfanz et al., 2007). From the KDE-converted KCCT scale scores, descriptive statistics were collected for each cohort to determine whether PP independently impacted movement from 8th grade performance of novice, apprentice, or proficient/distinguished to equal or higher levels of performance in the 11th grade.

In addition, paired-sample *t* tests determined the strength of the relationships between the 8th grade and their corresponding 11th grade test scores and compared the

differences between their means. Because the study lacked random assignment, we used a quasi-experimental, nonequivalent groups design (Shadish, Cook, & Campbell, 2002; Trochim & Donnelly, 2008) for a pretest-posttest comparison of KCCT math and a pretest-posttest comparison of KCCT social studies mean-centered scale scores for the non-PP and the PP cohorts. Because students did not experience PP in 8th or 11th grade social studies for either cohort, we studied social studies scores along with math scores for PP students to expose possible historical validity threats during the PP year. Equally improved scores in math and social studies for the PP cohort would reduce the validity of claims about the strength of PP's impact on student performance. Similarly, significant differences between math and social studies scores would suggest an impact by PP. KCCT 8th grade math and social studies tests served as pretests with their corresponding 11th grade assessments serving as posttests.

We performed an independent-samples *t* test for the equality of means (Hinkle, Wiersma, & Jurs, 2003) to determine the demographic comparability of the non-PP and PP cohorts on the following variables: minority membership, gender, free or reduced lunch (FRL) status, special education (ECE), and English as a second language (ESL) designation. We conducted a second independent-samples *t* test to compare the respective 8th grade pretest means of the KCCT math and social studies mean-centered scale scores. In addition to the validity threats of differences in demographics and initial academic performance inherent in non-equivalent group design, we considered historical and maturation factors (Trochim & Donnelly, 2008, p. 169). We conducted a one-way analysis of variance (ANOVA) (Hinkle et al., 2003) to determine if a statistically significant difference existed between cohorts in the mean KCCT math gains and social

studies gains from the 8th to 11th grade. This analysis employed mean math or social studies gain as a dependent variable and PP as the independent variable.

Results

Cohort Comparability

We did not randomly select the non-PP and PP cohorts; hence, this was a quasi-experimental design that began with tests to examine demographic and academic cohort comparability. A multivariate analysis of variance (MANOVA) (Cronk, 2010) was calculated to examine the effect of each cohort on the collection of demographic characteristics of minority, gender, FRL, ECE, and ESL. No significant effect was found on the collective demographic variables (*Wilks' Lambda*(5,499) = .99, $p > .05$), which prompted us to conduct subsequent univariate tests. Independent-samples *t* tests of non-PP and PP cohorts on minority, gender, FRL status, ECE, and ESL yielded significant differences, indicating demographic comparability of the cohorts (Table 21).

Table 21

Independent-samples t test Comparing Demographic and Pretest Variables

	Non-PP Cohort (N = 241)		PP Cohort (N = 264)		Equality of Means		
Demographic	*M*	*SD*	*M*	*SD*	*t*	*df*	*p*
Minority	.73	.44	.70	.46	.84	503	.40
Gender	.31	.47	.35	.48	-.89	503	.38
FRL	.78	.41	.80	.40	-.52	503	.60
ECE	.24	.43	.25	.44	-.34	503	.73
ESL	.02	.14	.05	.22	-1.76	458.89	.08
8th Grade KCCT							
Math	27.33	17.26	25.66	18.63	1.04	503	.30
Social Studies	29.53	14.55	26.27	16.26	2.38	502.80	.02

Note. FRL = Free or reduced price lunch. ECE = Exceptional Child Education (special education eligible). ESL = English as Second Language.

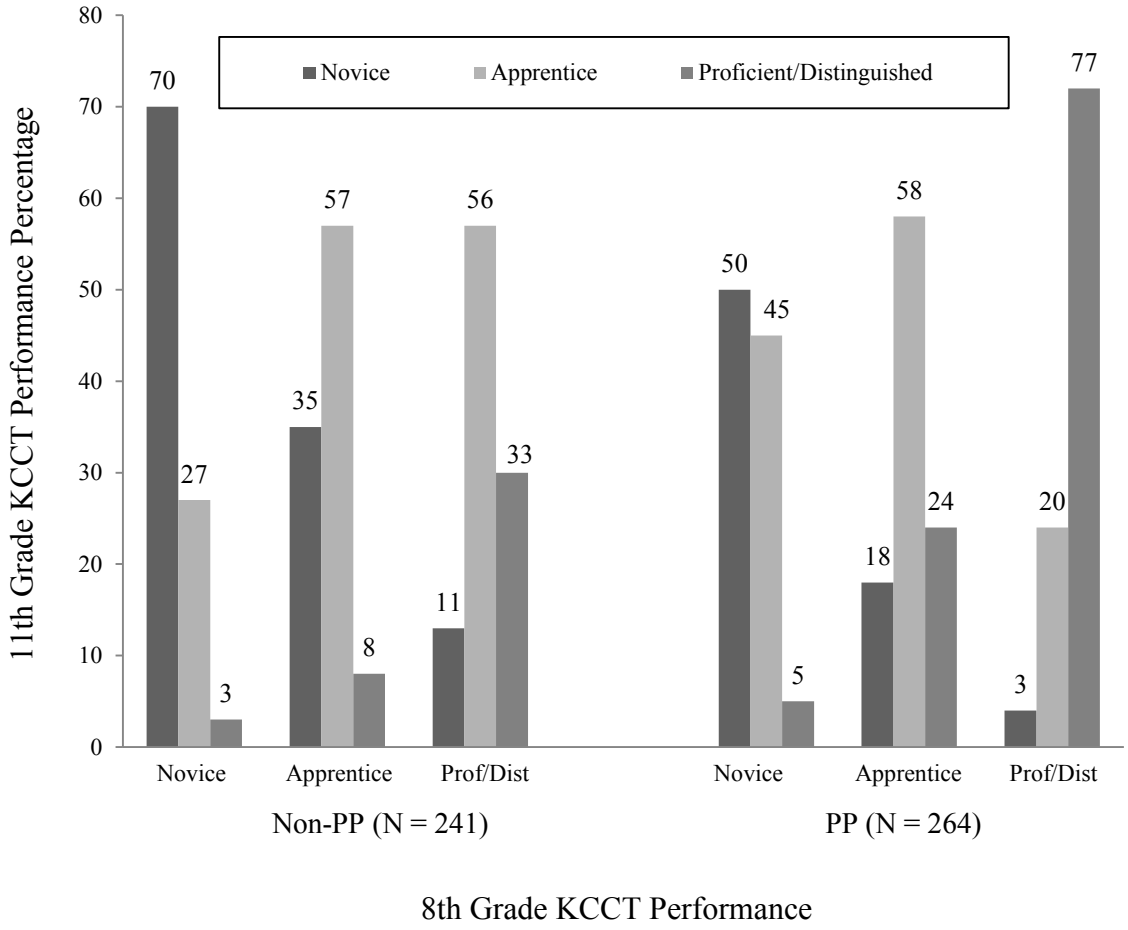
To control for initial academic differences between cohorts in math and social studies, we established the KCCT 8th grade scale scores as pretest covariates in an analysis of covariance (ANCOVA) (Hinkle et al., 2003). After controlling for differences in the 8th grade pretest scores, the ANCOVA indicated no significant differences in math means, and pairwise comparisons revealed a partial eta squared effect ($\eta^2 = .08$), which was medium to large (Cohen, 1988). The ANCOVA indicated a significant difference between 8th grade social studies pretest scores, but after controlling for initial differences, the analysis revealed no statistically significant differences in the variance in 11th grade social studies scores between the two cohorts ($F(1, 502) = 2.87$, $p > .05$). Corroborating the ANCOVA results for math, an independent-samples *t* test comparing the means of 8th grade KCCT math tests between non-PP and PP cohorts (Table 21) yielded no significant difference. The absence of significant differences between initial math and social studies pretests affirmed academic comparability of the two cohorts.

Changes in Performance Level

For math results, the PP at-risk cohort achieved considerable gains. We compared three initial performance levels of novice, apprentice, and proficient/distinguished for the non-PP and PP cohorts on their 8th grade tests with the subsequent percentages of at-risk students who moved from those initial levels to 11th grade levels of novice, apprentice, and proficient/distinguished (Figure 6). Descriptive statistics for math revealed that 14% of the PP students moved from 8th grade novice/apprentice to 11th grade proficient/distinguished on the KCCT math test as compared with 6% of the non-PP students. Also, approximately 77% of originally proficient/distinguished students in the

PP cohort repeated as proficient/distinguished on the more difficult 11th grade test as compared with 33% repeating the feat in the non-PP cohort.

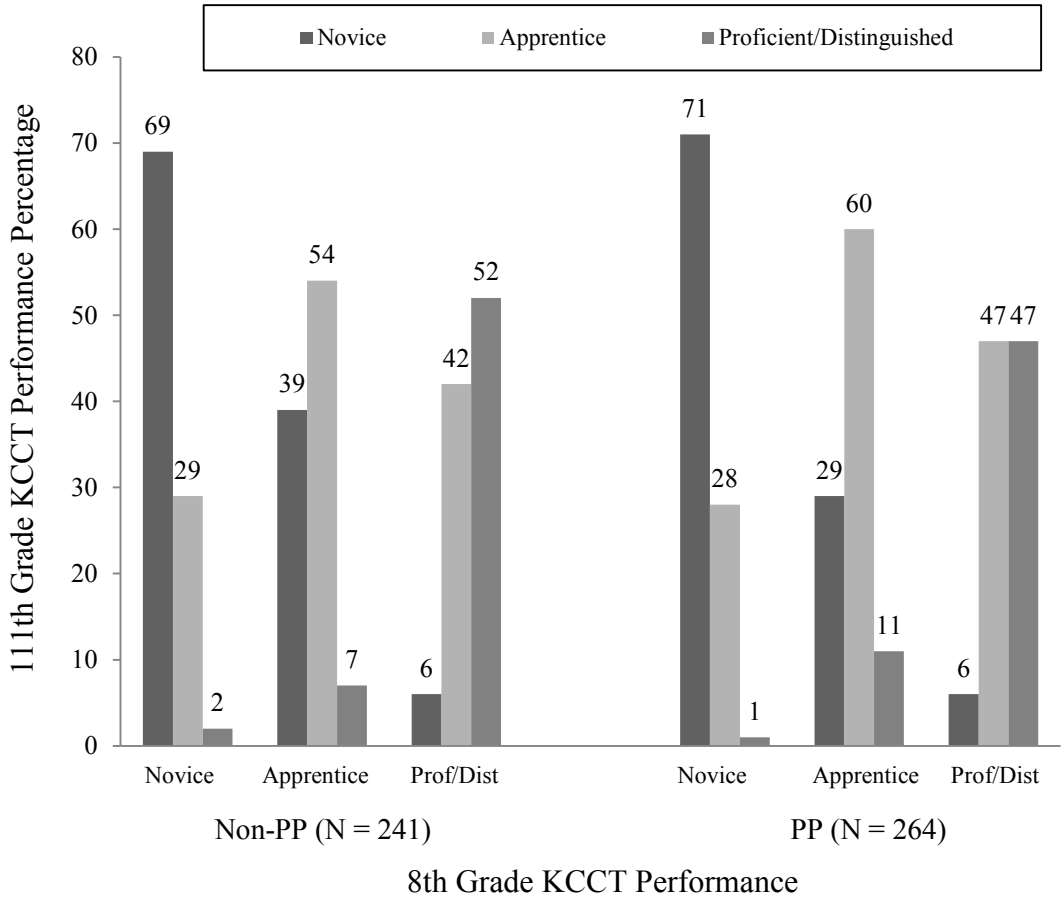
Figure 6. Performance Distribution of 8th Grade and 11th Grade KCCT Math Scores.



Similar comparisons revealed minimal movement for either cohort in social studies, a subject area that did not receive PP treatment (Figure 7). The PP cohort of 11th grade students who demonstrated a 44 percentage-point increase over their non-PP cohort counterparts repeating math proficient/distinguished performance actually produced 7% fewer students than the non-PP cohort maintaining that performance level in social studies. In general, after experiencing PP, distinct increases in math achievement by PP

cohort students occurred at every KCCT performance level.

Figure 7. Performance Distribution of 8th Grade and 11th Grade KCCT Social Studies Scores.



Comparisons of Mean-centered Scale Scores

For both cohorts, a Pearson correlation coefficient revealed a moderately strong positive relationship (Hinkle et al., 2003) between each pair of math and social studies assessments. Results confirmed that students with higher scores on 8th grade KCCT tests tended to also have higher scores on their 11th grade tests, regardless of subject area, which suggested comparability of the 8th and 11th grade tests. We performed paired-sample *t* tests to compare mean KCCT math and social studies scores from 8th to 11th grade (Table 22). The most prominent finding was the statistically significant increase by

the PP cohort from 8th to 11th grade math ($p < .001$) complemented by the practical significance of a small to medium effect size (Cohen, 1988) of .32 ($d = .32$). Results also indicated a significant increase for social studies for the PP cohort from 8th to 11th grade ($p < .05$). However, the effect size of .16 ($d = .16$) was in a range that Cohen (1988) identified as below small, and after controlling for a pretest covariate, the previously reported ANCOVA revealed no statistically significant differences in the variance in 11th grade social studies scores between the two cohorts.

Table 22

Comparison of Mean-centered Scale Scores for KCCT Math and Social Studies

	8th Grade		11th Grade		Descriptive Statistics		
Cohort/Subject	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Non-PP Cohort[a]							
Math	-11.56	17.25	-13.46	14.87	2.19	.03	.14
Social Studies	-11.46	14.56	-10.25	15.06	-1.48	.14	.10
PP Cohort[b]							
Math	-13.53	18.24	-8.13	17.81	-5.12	.00	.32
Social Studies	-13.99	16.32	-11.61	15.83	-2.55	.01	.16

Note. [a]N = 241. [b]N = 264.

In the PP cohort, in addition to a statistically significant increase in KCCT math scores from 8th to 11th grade, the average mean-centered score shifted from 13.53 points below the state mean in the 8th grade to 8.13 points below the state mean in the 11th grade, decreasing the gap by approximately 5.4 points. In contrast, for the non-PP cohort, the distance in the average mean-centered score from the state mean widened in math by approximately two points from 11.56 points to 13.46 points below the state mean. PP also impacted the relationship between prior achievement and future performance, as Pearson correlations revealed a decrease in the relationship between 8th

grade and 11th grade mathematics achievements, with values of $r = .66$ and .55 in PP and non-PP classrooms, respectively. Since prior achievement generally predicts future performance (Lee & Bryk, 1989; Raudenbush, 2004; Raudenbush & Willms, 1995), the reduced correlation strength between prior achievement and posttest math scores, combined with the increased mean, indicated that lower performing students on the 8th grade math test scored at higher levels on the 11th grade math test. Results suggested that PP impacted math proficiency among initially low-achieving students.

Quasi-experimental Analysis

We conducted a one-way ANOVA to compare the mean gains in KCCT math mean-centered scale scores of the two cohorts from 8th grade to 11th grade. Previous ANCOVA results revealing no significant differences in social studies scores eliminated the need for conducting an ANOVA for social studies. Corroborating the previous evidence of impact of PP on 11th grade KCCT math scores, we found a significant difference between cohort gain scores ($F(1, 503) = 49.42$, $p < .01$). Results indicated that students in the non-PP cohort decreased in their KCCT math gain score by nearly two points ($M = -1.79$, $SD = 13.4$) whereas the PP cohort increased by approximately seven points ($M = 7.05$, $SD = 14.74$). Due to the use of an ANOVA to compare two cohorts, we applied omega squared (Hinkle et al., 2003) to estimate the effect size, with $\omega^2 = .09$, which complemented statistical significance with practical significance. Borman, Hewes, Overman, and Brown (2003) concluded from a meta-analysis of comprehensive school reform that researchers can expect between a .09 and .15 effect size for district-wide samples for school reforms that "go beyond the effect of Title I" (p. 35).

Further strengthening the validity of the impact of PP, the means of the KCCT

11th grade social studies tests for the two cohorts were not significantly different even with initial pretest differences, but a significant difference was found between the means of the KCCT 11th grade math assessments ($t(499.10) = -4.35$, $p < .001$). Results suggested that PP significantly affected 11th grade math results since the PP cohort experienced the treatment in math and not social studies (Table 23).

Table 23

Independent-samples t tests Comparing Posttest Variables

	Non-PP Cohort (N = 241)		PP Cohort (N = 264)		Equality of Means		
11th Grade KCCT	*M*	*SD*	*M*	*SD*	*t*	*df*	*p*
Math	22.54	14.87	28.87	17.81	-4.35	499.1	.00
Social Studies	24.75	15.06	24.39	15.03	.26	503	.80

Discussion

This study indicated that when compared with a previous cohort of students with similar at-risk factors predictive of dropping out (Balfanz et al., 2007), PP students achieved statistically significant greater gains in the KCCT math scale scores between the 8th and 11th grades. After JCPS moved dramatic numbers of students in one year to proficient performance in math, we studied results for at-risk students to determine the comprehensive reach of PP. Based on findings for 11 PLA and near-PLA JCPS high schools that implemented PP during the 2010-2011 school year, achievement gains for students at-risk of dropping out suggested PP was a reform that moved to scale and demonstrated genuine, independent effects on student achievement at statistically significant levels.

The impact of PP on math performance for students at-risk of dropping out defied

the odds of a reform significantly increasing student achievement across an urban district (Zavadsky, 2009). Meaningful reform should increase the mean and decrease the variance in student achievement, which requires considerable academic movement of at-risk students. This study found that the mean math gain for at-risk students in the non-PP cohort actually decreased by nearly 2 points from the 8th to the 11th grade, whereas the PP cohort increased by nearly 5 points, a statistically significant and educationally meaningful reversal. In addition, PP cohort students notably improved their performance from 8th to 11th grade. Even if they maintained their performance levels according to nominal state classification, on the KCCT math tests from 2007 to 2011, an annual average of 17% fewer Kentucky students scored proficient/distinguished on the 11th grade tests than they did on their corresponding 8th grade tests (Kentucky Department of Education, 2011), suggesting differences in difficulty level, rubrics, and cut scores on middle and high school math assessments. Despite the possibly higher difficulty level of the high school math assessment, 77% of at-risk students who experienced PP repeated as proficient or distinguished from 8th to 11th grade. This was nearly 44 percentage-points higher than those who maintained that performance level in the non-PP cohort.

To complement the established external validity of the KCCT state assessments, the internal validity of this study was strengthened by demonstrating no statistically significant demographic or academic differences between the non-PP and PP cohorts of at-risk students. We only included at-risk students who had not dropped out, had arrived without delay to the 11th grade, and had corresponding 8th and 11th grade test scores, arguably excluding some of each cohort's most struggling students. Nevertheless, although we excluded some of the at-risk students, the identified sample represented a

sizeable component of the current generation of students who have annually sat in front of us, at-risk of dropping out (Balfanz et al., 2007), devoid of effective reform interventions, and plagued with historically low achievement scores characteristic of PLA urban high schools.

To further strengthen the non-equivalent group design, social studies scores were examined alongside math scores for PP cohort students. Given that social studies was not connected with PP and had the least similarities with math content, significant increases in both math and social studies could rule out a strong relationship between PP and math gains. However, for the PP cohort, the same students taking a different test, we found no significant difference in social studies gains even when controlling for an 8^{th} grade pretest covariate. Therefore, statistically significant achievement gains, considerable increases in proficient performance of at-risk students across the district, and a noteworthy effect size strengthened the credibility of PP as an effective high school reform for at-risk students.

As with all research and implications, this study contains limitations that potentially temper the findings. First, as outlined by the PP guidelines, the study assumed that, on average, the PLA schools' math teachers implemented district curriculum guides and standards, evaluated student competency through standards-based evaluation of work, sought to guarantee student competency of key standards each grading period, and ensured a fail-safe score on each grading-period summative assessment. Second, the study examined the effects of a single-year treatment. A multi-year study would provide time for trend data. Third, although the non-PP and PP cohorts proved statistically comparable, PP was implemented amid the historical threat of state

sanctions and staff removal, which could have produced reactionary emotions of fear or anger for students and staff members. Yet, social studies teachers did not increase scores. Fourth, although each student in both cohorts met the dropout-predictive criteria, the study focused on a specific population, including only those who made it to the 11th grade on time, not including those who dropped out or fell behind. Due to the unavailability of how many students from each cohort had dropped out of school by the fall of 11th grade (pre-PP) and during the 11th grade (during PP vs. non-PP), future studies could employ survival analysis statistics (Singer & Willett, 1993) that do not totally require complete cases such as corresponding 8th and 11th grade test scores for each participant. Finally, all but one of the 11 schools in the study received some level of state assistance that may have influenced achievement gains. Four schools received state-provided resource teachers and substantial financial resources for professional development and stipends for extended staff time, and six schools received moderate supplementary funding. However, it is more likely that such confounds slightly mitigate rather than negate PP as a scalable high school reform given the statistically significant growth across 11 schools in math and the lack of significant change in social studies.

Implications for Practice, Policy, and Research

Establishing effective high school reform that rapidly moves a significant number of students to levels of proficiency across an urban district continues to confront educational practice, policy, and research. Darling-Hammond et al. (2005) affirmed, "high schools have presented a perennial challenge to school reform efforts" (p.169). Although JCPS high schools realized significant gains for at-risk students in reading and math after implementing PP in 2010-2011, most of the PLA schools have remained

ranked in the bottom performing schools in Kentucky. In 2011-2012, JCPS high school teachers in PLA and near-PLA high schools entered the second year of PP implementation facing new standards, completely different state assessments, six principal changes, and considerable restaffing of faculties. However, the powerful effect of the previous year's results and productive instructional practice compelled schools to maintain the momentum of PP and the goal to guarantee competency of key standards for every student. Ensuring learning had provided purpose and instructional coherence across the high school level, created conditions for ownership and collective efficacy of staff and students, and generated hope for the possibility of accelerated effective reform across an urban district. The principles of PP could permanently change JCPS, and they deserve further research and enhancement for the benefit of our students most at-risk of dropping out.

Guided by the overarching PP goal to guarantee key-standard competency for each student, districts and administrators should adopt the most reproducible elements of PP: to create conditions of urgency (Stringfield, Reynolds, & Schaffer, 2008), instructional coherence of curriculum, instruction, and assessment (Oxley, 2008), and "co-construction" (Stringfield & Datnow, 2002, p. 269) opportunities with teachers for implementation and decision-making. Teachers should claim the power of collectively ensuring student learning by collaboratively evaluating student understanding of standards, instead of settling for averaging grades (Guskey, 2009; Lekholm & Cliffordson, 2008; Marzano, 2010). They should also create common formative assessments to measure individual student progress, engage learners in self-reflection, and seek instructional implications (Stiggins & DuFour, 2009). Finally, teachers should

adjust instruction and interventions to guide each student to demonstrate an acceptable level of competency in key standards.

The results of our study also indicated that legislators and school boards should provide fewer prioritized goals, invest in existing principals and teachers, and support systems and processes. Amid mandates to teach a growing number of new standards, PP provided practitioners with only three prioritized standards per six-weeks grading period for ensured student competency. Schools were provided common diagnostic and summative assessments on fewer standards, a goal each grading period to guarantee a level of individual student competence that teachers respected, and the balance of supervision and support for general implementation. Policymakers should consider these PP practices and "enable schools to enhance the stability and professional capacity of staff members and the academic performance and active engagement of students" (Berman & Camins, 2011, p. 28).

Sustainability of PP beyond a one-year significant result provides a challenge for researchers. Hargreaves and Fink (2006) proposed that after effectively implementing change, "the biggest challenge of all is to make it durable and sustainable" (p. 2). The principles of Highly Reliable Organizations (HRO) (Stringfield et al., 2010), implemented since 1996 in the extraordinarily successful and sustained reform effort in economically deprived southern Wales, could provide researchers with a framework for examining the potential for PP sustainability. The PP principle of guaranteed competency parallels the HRO precept that failure would be disastrous. In addition, PP real-time tracking of student competency of fewer key standards matches HRO fundamentals of clear goals and powerful databases. Further, PP collaborative teams of

practitioners who collectively construct ways to ensure individual student learning mirror the HRO design to be "tight on specifying concepts and systems to be used, and loose on organizational processes and details determined by schools" (Stringfield et al., 2010, p. 27).

Evidence supports that through PP, JCPS educators and students in 11 high schools moved high school reform to scale at the district level for students most at-risk of dropping out. Although results were significant, these schools have only begun their journey to move each student to proficient performance. To sustain and build from these gains, they must depend on additional expertise from educational peers, policymakers, and researchers to help them maintain their momentum of urban high school reform. Fellow practitioners must implement PP principles and add to the collective knowledge base about standards-based grading, collaborative formative work, and ensured student learning of key standards. Policymakers should create conditions for co-construction of reform details, and researchers should study the sustainability of PP as a scalable high school reform. Through PP, JCPS educators achieved very significant gains with its high school at-risk students, and this study provides evidence, traction, and hope for understanding elusive urban high school reform.

EXECUTIVE SUMMARY

Faced with the threat of state and federal sanctions and the desire to move unprecedented numbers of students to proficiency, PP challenged district and school leaders in 11 Jefferson County, Kentucky, high schools to focus on three key standard categories in reading and mathematics classes during a grading period, create formative tasks through which students demonstrate competency for each standard, guide each student to a level of competency in each standard prior to the end of a six-week grading period, and create fail-safe assessments that enable students to recover missed Algebra 2 content. In three quantitative studies, we examined the relationship between grades and student achievement, controlled for classroom and school compositional effects to examine within- and between-class variation in achievement, and investigated the reform's impact on a large urban school district's students most at-risk for dropping out. The results revealed that PP positively impacted student achievement and contains the conditions necessary as a scalable urban high school reform design.

Project Proficiency's Impact on Student Achievement

Quantitative analysis using OLS regression, HLM, and analysis of variance determined that PP strengthened the relationship between grades and performance on state mathematics assessments, improved student achievement, and provided an effective support to students most at-risk for failure and subsequent drop out. First, the study revealed that PP accentuated the association between classroom grades and academic

achievement. Within both the PP cohort and the non-PP cohort, grades had a low-positive association with KCCT test scores. For students evaluated on a standards-based grading approach, the association between grades and test scores was stronger than those students evaluated on a traditional grading model. All students, minority students, and at-risk students had stronger correlations between grades and state assessments when experiencing standards-based grading.

Most importantly for the research on standards-based grading within PP, grades became much more of a valid indicator of achievement as measured by success on the KCCT assessment. Students who experienced traditional grading methods in both cohorts scored below proficient nearly 75% of the time even though they received an A or a B in the specific content class. For students experiencing standards-based grading, over 55% of students scored above proficient when they received an A or a B in the content course. As a result, over twice as many students scored proficient or above on the state assessment when successfully scoring above average on a standards-based grading approach as opposed to a traditional grading model. We determined that standards-based grades are a more valid and reliable predictor of student achievement than traditional-based grades.

Second, through HLM models controlling for prior achievement and SES at the student level and SES and PP implementation at the classroom level, PP increased mathematics achievement and decreased variation between classrooms. In PP classrooms, mathematics scale scores increased nearly one-half of a performance level on the state assessment and yielded a 22% increase in students reaching state-established proficiency benchmarks. We found a statistically significant decrease in between-

classroom variation in PP classrooms, with estimates diminishing from 31% to 14% under PP. We concluded that PP ameliorated the negative effect of classroom SES on student achievement and the combination of improved mean achievement and decreased variation between classes implies that instructional practices changed with large numbers of teachers across PP schools.

Third, empirical evidence strongly suggested that PP impacted the math achievement of students most at risk of dropping out of school. We found statistically significant increases in mathematics achievement for at-risk PP students who met dropout-predictive criteria (Balfanz et al., 2007). The study revealed that 14% of PP students who scored below proficient on the 8th grade KCCT met the proficiency benchmark in the 11th grade, as opposed to 6% in the non-PP group. In addition, 77% of PP students with proficient or higher 8th grade results scored proficient in 11th grade, as opposed to only 33% in the non-PP group. Finally, statistically significant gains were revealed in KCCT math scale scores from the 8th grade to the 11th grade.

Project Proficiency as a Scalable Instructional Design

Faced with external demands to rival international academic standards, produce a globally competitive workforce, and rapidly move dramatic numbers of students to levels of proficiency, educational practitioners, policymakers, and researchers continue to search for effective urban high school reform scalable at the district level. In this capstone, results of three empirical studies of the impact of PP across 11 high schools in a large urban district indicated that teachers more accurately evaluated student work, classrooms more equally provided instructional quality, and at-risk students significantly increased their achievement in mathematics. We suggest that the PP design not only

reached every JCPS PLA high school, but also exhibited the "multidimensional nature" (Coburn, 2003, p. 3) of authentic scalability through its "depth, spread, shift of ownership, and sustainability" (p. 4).

PP consisted of four facets. First, the district provided a curriculum map, diagnostic assessment, and end-of-six-weeks summative proficiency assessment for each grading period based on three key standards or content categories. Second, teachers were asked to create tasks and assignments through which students could demonstrate competency for each key standard. Competency was not proficiency or mastery, but a level of understanding that a teacher respected. Teachers evaluated student work with standards-based grading, evaluating student understanding rather than simply averaging numbers. Third, with the help of district resource teachers, curriculum specialists, and local administrators, teachers collaboratively and collectively sought to guide 100% of their students to demonstrate competency for each of the three key standards by the time the proficiency assessment was administered. The district provided a web-based, electronic system for teachers to collect and track results of diagnostic, formative, and summative assessments. Fourth, students scoring below 80% on a six-weeks proficiency assessment were guided to recover missed content until they scored 80% or higher.

Depth. The results of this study revealed that the interrelated dimensions of scale (Coburn, 2003) were evident in each of the four facets of the PP reform. The author noted that a depth in the change of classroom practice is necessary to bring substantial educational reform to scale. According to Coburn (2003), "Because teachers draw on their prior knowledge, beliefs, and experiences to interpret and enact reforms, they are likely to gravitate towards approaches that are congruent with their prior practices, focus

on surface manifestations, rather than deep pedagogical principles" (Coburn, 2003, p. 4). The change to a standards-based grading approach within PP forced teachers to make profound changes to pedagogical principles that had dominated their classroom practice for decades. The four main tenets of PP aided teachers in their transition to a standards-based grading approach and led to "deep and consequential change in classroom practice" (Coburn, 2003, p. 4).

Based on the data from this research, the depth of instructional change caused by a move to standards-based grading led to an increased association between grades and achievement scores. The four main tenets of PP created conditions for depth in instructional change. First, math curriculum was condensed to three key standards for each six weeks. Prior to PP, the math curriculum was driven by state core content and district pacing guides that led to coverage of many topics with little emphasis on mastery of key standards. In order to implement a standards-based grading approach, PP established three key standards for each grading period. PP enabled teachers to evaluate students on their attainment of the three key standards for each six weeks.

Second, teachers created tasks that would demonstrate competency of the three key standards. Teachers could no longer use a hodge-podge of factors to determine a students' grade for the six weeks period. Every task or assignment given to students directly measured their competency in one of the key standards for the six weeks. At the conclusion of each grading period, students were evaluated solely on their attainment of the three key standards established by PP. This move to standards-based grading was the change in pedagogical principles that established the depth required to move educational reform to scale.

Third, and possibly most important, teachers guaranteed the competency of every student in the three key standards. Teachers had to collaboratively find ways to establish interventions for students who did not attain proficiency in any of the key standards. Diagnostic and formative assessments provided data for teachers on student mastery of the three key standards for the grading period. Without the pedagogical change to standards-based grading, guaranteed competency of the standards would be impossible. Teachers had to evaluate the achievement level of students through standards-based grading in order to guarantee their competency in the key standards.

Fourth, PP provided a fail-safe opportunity for students to recover any standard that they had not mastered at the conclusion of the grading period. If the student still had not met proficiency standards by this time, teachers would provide remediation opportunities in order to guarantee competency. By establishing a depth of instructional change within all schools through implementation of standards-based grading, teachers were able to evaluate student progress on the three key standards, grade students on their attainment of those standards, and provide opportunities after the grading period to recover missed standards and improve student grades. Through the implementation of standards-based grading in these schools, the four facets of PP provided the depth necessary to bring this instructional reform to scale.

Spread. PP created conditions for spread between classrooms and across 11 high schools in Jefferson County. Coburn's (2003) scalability framework identified the need for schools to "move beyond the spread of activities, materials, and structures to the propagation of 'beliefs, norms, and principles'" (p. 7). Through guaranteeing competency in three key standards, PP challenged teachers and school leaders to change

existing beliefs about teaching and learning. In the new context, all students were required to reach a degree of competency prior to taking an end-of-unit proficiency assessment, thus requiring teachers to own the achievement results for each of their students. As opposed to prior practices that assigned low grades for students not demonstrating competency and the continuing with course content, PP required teachers and their PLCs to create instructional activities and targeted interventions to ensure competency prior to the end of a six-week grading period. The shared responsibility for students learning created conditions for substantive collaboration among teachers.

The combination of new instructional practices and supporting structures reflected a change in beliefs in PP schools. Professional collaboration became a new and highly-valued norm for PP teachers. During formal and structured PLC meetings, teams of teachers met to: review learning targets and plan instructional activities; examine student work to identify trends, needs, and instructional strategies; review diagnostic, formative, and proficiency data; and plan for school-wide student interventions. This increase in collective efficacy represents the spread of deep changes in instructional practices across classrooms and schools in Jefferson County. We find additional support for this finding. After controlling for prior achievement and individual and classroom SES, PP implementation decreased variation among classrooms by approximately 55%. Given the fact that PP was implemented in 110 classrooms in 11 schools, reductions in between-classroom variation from 31% to 14% suggests that PP's changes in instruction, norms, beliefs, and principles created conditions for the spread of depth across JCPS classrooms. Carroll (2009) observed the futility of individual teachers working alone and their inability to "know and do everything to meet the needs of 30 diverse students every day

throughout the school year" (p. 10). The author called for schools to become places that value and take full advantage of teamwork that is part of high-performance organizations across the world. Through implementing the four key elements of PP, strong collaboration created spread among teachers, ensured competency for all students, and deepened the instructional knowledge base.

In addition to identified increases in the number of schools implementing a reform, Coburn (2003) identified spread as to what extent district policies, procedures, and professional development reflect a reform effort. Jefferson County leaders challenged schools to examine traditional approaches to teaching and create a new instructional approach that "guarantees competency" for all students through an emphasis on key standards and changes in assessments, interventions, and support mechanisms. As opposed to the traditional instructional planning activities where individual teachers and schools review core content curriculum documents and create learning objectives and supporting activities, PP shifted the curriculum alignment responsibilities to the district and gave teachers a clearly-articulated and aligned set of learning standard categories for each grading period. District leaders promoted the spread of the PP design across schools by creating diagnostic and proficiency assessments aligned to three key standards in the PP design and creating district-wide professional development on formative assessment, instructional strategies, and professional collaboration.

PP was a unified approach to learning that changed pedagogical and philosophical beliefs and reflected the "normative coherence" that Coburn (2003) identified as a necessary element to reach scale. With 11 schools and 110 classrooms participating in PP, district officials operated as a strategic agent charged with creating curriculum and

support materials for teachers in addition to allocating existing technological and technical assistance resources to support the effort. Furthermore, by providing opportunities for teachers and school leaders to participate in PP-specific professional development, district leaders deepened an individual school's capacity to implement the reform design. With results from the study reporting that PP increased mean achievement, strengthened the association between classroom grades and achievement, reduced variation between classrooms across the district, and improved achievement for the districts most at-risk students, we propose that the elements of PP spread across classrooms and schools in Jefferson County.

Shift of ownership. By bringing coherence to the four facets of PP, JCPS created conditions for the shift of ownership of reform from the district to teachers and students. To genuinely move to scale, Coburn (2003) asserted that a reform must shift from external to internal "authority for the reform held by those who have the capacity to sustain, spread, and deepen reform principles" (p. 7). Although a literature review revealed that effective urban high school reform for at-risk students included early middle school intervention, whole-school high school reform, and a supportive learning environment for the current generation of students in the desks in front of us, these reforms were too late, too slow, or too little to immediately impact significant numbers of at-risk students. However, without the luxury of previous interventions or a prescribed reform provided by the district, JCPS teachers in 11 PLA high schools significantly increased math achievement by at-risk students who experienced PP.

Through a "reconceptualization of proprietorship" (McLaughlin & Mitra, 2001, p. 317), PP shifted from an external to an internal reform by narrowing the state's growing

number of content standards to three key standards each grading period, providing corresponding diagnostic and summative assessments, and holding schools accountable for results rather than activity. Teachers were recruited to help identify the standards, create the assessments, and own the design. Once schools received the key standards and assessments, local administrators and teachers determined the sub-content and learning targets they believed best prepared students to understand each key standard and aligned lessons with the learning targets. Each school was allowed the flexibility to determine its own learning targets and was accountable for summative assessment results rather than adherence to prescribed sub-content. When principals led course-common learning teams of teachers to compare assessment results, and district officials guided teams of principals to collaboratively examine summative test scores, practitioners borrowed and exchanged lessons and learning targets and increasingly constructed, owned, and helped spread the most effective curriculum.

PP shifted ownership of instructional development by creating conditions for teachers to evaluate student competency using standards-based grading of assignments and tasks. "Adapting to local contextual needs" (Datnow & Stringfield, 2000, p. 195), teachers were free to define what competency looked like in student work and the district provided a web-based system for recording individual student competency for each key standard. The web-based program lifted the burden of grading from teachers by ultimately converting competencies into daily grades for students. The knowledge required for PP reform rested with the practitioners who were allowed to create student work through which students demonstrated understanding, design lessons that prepared students for those tasks, and evaluate whether students met a level of acceptable

competency through the tasks. Scoring student work and averaging grades using a point system require a good calculator and at least a teacher's aide, but evaluating student understanding for competency demanded decision-making and ownership by the instructor.

By establishing the PP goal of guaranteed competency of key standards each grading period, JCPS shifted ownership of assessment to practitioners and students. In addition to acquiring responsibility for establishing learning targets, designing lessons, and evaluating student tasks for understanding, teachers assumed responsibility for student learning and demonstration of competency before the end-of-six-weeks summative assessment. Teachers moved from independence and isolation in their classrooms to dependence on one another and eventual interdependence to collectively reinvent their instruction, assessment, and intervention practices (Allensworth & Easton, 2007). Administrators provided and facilitated learning team opportunities for teachers through creating common planning time and "mechanisms for ongoing learning" (Coburn, 2003, p. 8) as opposed to the usual checklists, required meetings, and completion of compliance documents. Teachers owned student results, relied on one another's expertise, and developed an unprecedented collective efficacy in JCPS high schools.

In addition, the PP goal of guaranteed competency elicited student ownership of learning. The district strongly suggested that student reflection count for 20% of the final grade each grading period. Teachers were allowed to collaboratively design means for students to reflect about progress toward competency and misunderstandings of standards. Although reflection designs varied across the PP schools, after the diagnostic

assessments and daily formative assignments, students were guided at every site to own their own learning by describing where they were on learning continua toward competencies, requesting assistance by specific standard, and realizing that below standard work meant "not yet" instead of failure. In fact, due to the PP fail-safe requirement that students must eventually score 80% or higher on each six-weeks summative assessment, teachers and students collaboratively discussed ways to move from remediation to recovery of competency for each key standard, building a "sense of community that empowered students" (Wilkins, 2008). PP shifted ownership of assessment to teachers and students and created conditions for expected, possible, probable, and inevitable learning for students most at-risk of dropping out.

Sustainability. Although we examined the impact of PP after only one year of reform, educational practitioners, policymakers, and researchers should consider the parallels of PP with HRO principles as evidence for the sustainability of PP. Fink and Hargreaves (2006) concluded that secondary school reforms were typically unsustainable. However, since 1996, the Neath-Port Talbot (NTE) Local Education Authority in an economically disadvantaged area in southern Wales, Great Britain, has sustained its implementation of HRO principles, and after equaling the Welsh national average in 2000, moved its test scores in 2007 considerably above the national average (Stringfield et al., 2010). Confidence for the sustainability of PP lies in its alignment with many of the customized and sustained HRO principles implemented by NTE including the urgency to succeed, a finite set of shared goals, powerful databases, a balance of tight and loose standard operation procedures (SOP), and collegial decision making (Stringfield et al., 2008; Stringfield et al., 2010).

The PP goal to guarantee competency of key standards by each student produced an HRO-like urgency unlike previously implemented high school reforms in JCPS, and aligning district curriculum maps, common assessments, and intervention supports with three key standards each grading period matched the HRO principle of establishing a clear and finite set of goals. The JCPS web-based tool for tracking diagnostic, formative, and summative assessment data corresponded to the HRO practice of gathering and effectively using data. The JCPS tight expectations for common key standards, standards-based grading, and ensured learning, balanced by looseness of demands for processes and local implementation details reflected the HRO recommendations for complimentary district-school SOP. Finally, the PP shift of ownership for decision-making about processes to improve curriculum, instruction, assessment, resources, and SOP affirm the most convincing parallel with HRO principles, allowing for continuous improvements from practitioners who are "flaw finders and process/program improvers" (Stringfield et al., 2010, p. 15).

Results of this study of PLA high schools in a large urban district indicated a strong relationship between PP and teacher grading practices, classroom impact, and student achievement of students at-risk of dropping out. We found significant results across all 11 high schools that implemented PP, and the reform meets the scalability litmus test for depth, spread, shift of ownership, and sustainability. PP principles required no additional money, staff, or purchased programs and relied on "strong systems rather than strong or unusually effective people" (Reynolds, Creemers, Stringfield, Teddlie, & Schaffer, 2002, p. 289), increasing its probability for scalability and sustainability. Although a combination of factors influenced the achievement gains

associated with PP, the results of this study provide hope and demand for further research for PP's potential to establish high school reform to scale in an urban district.

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.

Abbott, S. E., & Fisher, P. D. (2011). *Harnessing teacher knowledge. A guide to developing school-based systems for professional learning and planning.* Retrieved from Great Schools Partnership website: http://www.greatschoolspartnership.org/pdf/HTK_Tool_Print.pdf

Aiken, W. M. (1942). *The story of the eight year study.* New York, NY: Harper.

Alexander, K.L., Entwisle, D.R., & Kabbini, N.S. (2001). The dropout process in life course perspective: Early risk factors at home and school. *Teachers College Record*, *103*(5), 760-882.

Allen, D., & Blythe, T. (2004). *The facilitator's book of questions: Tools for looking together at student and teacher work.* New York, NY: Teacher's College Press.

Allen, J. D. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House*, *78*(5), 218-223.

Allensworth, E. M., & Easton, J. Q. (2005).*The on-track indicator as a predictor of high school graduation*. Consortium on Chicago School Research at the University of Chicago. Retrieved from http://www.ccsr.uchicago.edu/publications/p78.pdf

Allensworth, E. M., & Easton, J. Q. (2007). *What matters for staying on-track indicator and graduating in Chicago public high schools*. Chicago: Consortium on Chicago School Research at the University of Chicago. Retrieved from http://ccsr.uchicago .edu/publications/0%20What%20Matters%20Final.pdf?q=graduating-students-research-ready

Almeida, C., Steinberg, A., Santos, J., & Le, C. (2010). *Six pillars of effective dropout prevention and recovery: An assessment of current state policy and how to improve it*. Jobs for the Future. Retrieved from http://www.jff.org/sites/default/ files/DropoutBrief-090810.pdf

America's Promise Alliance (2010). Retrieved from http://www.americaspromise.org /About-the-Alliance.aspx

American Institutes for Research (1999). *An educator's guide to school-wide reform.* Washington, DC: Author.

American Institutes for Research (2006). *CSCQ center on middle and high school comprehensive school reform models.* Washington, DC: Author.

Atherton J. S. (2011). *Teaching and learning: What works and what doesn't*. Retrieved from http://www.learningandteaching.info/teaching/what_works.

Aud, S., Hussar, W., Planty, M., Snyder, T., Bianco, K., Fox, M.,…Drake, L. (2010). *The Condition of Education 2010* (NCES Report 2010-028). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Austin, G. R. (1979). Exemplary schools and the search for effectiveness. *Educational Leadership, 37*(1), 10-12.

Baker, J. A. (2006). Contributions of teacher–child relationships to positive school adjustment during elementary school. *Journal of School Psychology, 44*(3),211-229.

Balfanz, R., Herzog, L., & MacIver, D. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist, 42*(4), 223-235.

Balfanz, R., & Legters, N. (2004a). Locating the dropout crisis: Which high schools produce the nation's dropouts? In G. Orfield (Ed.), *Dropouts in America: Confronting the graduation rate crisis* (pp. 57-84). Cambridge, MA: Harvard Education Press.

Balfanz, R., & Legters, N. (2004b). *Locating the dropout crisis: Which high schools produce the Nation's dropouts? Where are they located? Who attends them?* (Research Report No. 70). Washington, DC: Center for Social Organization of Schools, Johns Hopkins University.

Balfanz, R., Legters, N., West, T. C., & Weber, L. M. (2007). Are NCLB's measures, incentives, and improvement strategies the right ones for the nation's low-performing high schools? *American Educational Research Journal 44*(3), 559-593.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Sciences, 29*(1), 37-65.

Barber, B. K., & Olsen, J. A. (2004). Assessing the transitions to middle and high school. *Journal of Adolescent Research, 19*(1), 3-30.

Berends, M., Bodilly, S., & Kirby, S. N. (2002). Looking back over a decade of whole-school reform: The experience of new American schools. *Phi Delta Kappan, 84*(2), 168-175.

Berends, M. & Peñdoza, R., (2010). Increasing racial isolation and test score gaps in mathematics: A 30-year perspective. *Teachers College Record, 112*(4), 978-1007.

Berman, S. H., & Camins, A. (2011). Investing in turnaround that endures. *Education Week, 31*(10), p. 28.

Bilko, R., & Ladd, H. F. (2006). The impacts of charter schools on student achievement: Evidence from North Carolina. *Education Finance and Policy, 1*(1), 50-90.

Bodilly, S. J., Glennan, T. K., Kerr, K. A., & Galegher, J. R., (2004). *Expanding the reach of education reforms: Perspectives from leaders in the scale-up of educational interventions.* Santa Monica, CA: Rand Corporation.

Bodkin-Andrews, G. H., O'Rourke, V., Dillon, A., Craven, R. G., & Yeung, A. S. (2009). *Explaining away aboriginality: Causal modeling of academic self-concept and disengagement for indigenous and non-indigenous Australian students*. Paper presented at the Fifth Global SELF International Biennial Conference, United Arab Emirates University, Al Ain.

Borman, G. D., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's equality of educational opportunity data. *Teachers College Record, 112*(5), 1201-1246.

Borman, G. D., Hewes, G., Overman, L. T., & Brown, S. (2003). Comprehensive School Reform and Achievement: A Meta-Analysis. *Review of Educational Research, 73*(2), 125-230.

Bowers, A. J., (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration*, *47*(5), 609-629.

Brand, B. (2009). *High school career academies: A 40-year proven model for improving college and career readiness.* Commissioned by the National Career Academy Coalition. Retrieved from http://www.aypf.org/documents /092409CareerAcademiesPolicyPaper.pdf

Brennan, R. T., Kim, T., Wenz-Gross, M., & Sipperstien, G. N. (2001). The relative equitability of high-stakes testing versus teacher assigned grades: An analysis of Massachusetts Comprehensive Assessment System. *Harvard Education Review, 71*(2), 173-216.

Bridgeland, J., Dilulio, J., & Morrison, K. (2006). *The silent epidemic: Perspectives of high school dropouts*. Washington, DC: Civic Enterprises, LLC.

Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice*, *10*(1), 35-36.

Brookhart, S. M. (1993). Teachers grading practices: Meaning and values. *Journal of Education Measurement, 30*(2), 123-142.

Brookhart, S. M. (1994). Teachers grading: Practice and theory. *Applied Measurement in Education, 7*(4) 279-301.

Brophy, J. (1988). Research on teacher effects: Uses and abuses. *The Elementary School Journal, 89*(1), 3-21.

Brophy, J. & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3d ed., pp. 328-375). New York, NY: Macmillan.

Brown-Jeffy, S. (2009). School effects: Examining the race gap in mathematics achievement. *Journal of African American Studies, 13*(4), 388-405.

Bruce, M., Bridgeland, J. M., Fox, J. H., & Balfanz, R. (2011). *On track for success: The use of early warning indicator and intervention systems to build a grad nation*. Washington, DC: Civic Enterprises, Johns Hopkins University.

Bryan, K., Klein, D., & Elias, M. J. (2007). Applying organizational theories to action research in community settings: A case study in urban schools. *Journal of Community Psychology*, *35*(3), 383-398.

Bryk, A. S. (2009). Support a science of performance improvement. *Phi Delta Kappan, 90*(8), 597-600.

Bryk, A. S. (2010). Organizing schools for improvement. *Phi Delta Kappan, (91)*7, 23-28.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.

Bryk, A. S., Sebring, P. B., Kerbow, D., Rollow, S., & Easton, Q. (1998). *Charting Chicago school reform.* Boulder, CO: Westview Press.

Carroll, T. (2009). The next generation of learning teams. *Phi Delta Kappan, 91*(2), 8-13.

Childress, S., Elmore, R., & Grossman, A. (2006). How to manage urban school districts. *Harvard Business Review, 84*(11), 55-68.

Childress, S., Elmore, R., Grossman, A., & Akinola, M. (2004). Note on the PELP coherence framework. Public Education Leadership Project at Harvard University. Retrieved from http://www.kasa.org/professionaldevelopment /documents/PELPFramework.pdf

Cizek, G.J., Fitgerald, S. M. & Rachor, R. E. (1996). Teachers assessment practices: Preparation, isolation and the kitchen sink. *Educational Assessment*, *3*(2), 159-179.

Coalition of Essential Schools (2011). *The CES Common Principles*. Retrieved from http://www.essentialschools.org/items/4

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, *32*(6), 3-12.

Cohen, J. (1988). *Statisical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: Government Printing Office.

Conley, D. T. (2000, April). *Who is proficient: The relationship between proficiency scores and grades*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Consortium on Chicago School Research (2010). C*hicago high school redesign initiative: Schools, students, and outcomes.* Chicago, IL: Author.

Crandall, D. P., Loucks-Horsley, S., Bauchner, J.E., Schmidt, W.B., Eiseman, J.W., & Cox, P. L. (1982). *People, policies and practices: Examining the chain of school improvement* (Vols. 1–10). Andover, MA: The Network.

Creswell, J. W., & Plano-Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Cronk, B. C. (2010). *How to use PASW statistics: A step-by-step guide to analysis and interpretation* (6th ed.). Los Angeles, CA: Pyrczak Publishing.

Cross, C.H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53-72.

Cuban, L. (1984). *How teachers taught: Constancy and change in American classrooms, 1890-1980.* New York, NY: Longman.

Darling-Hammond, L. (1997). Toward what end? The evaluation of student learning for the improvement of teaching. In J. Millman (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 248-263). Thousand Oaks, CA: Corwin.

Darling-Hammond, L. (2006). No Child Left Behind and high school reform. *Harvard Educational Review, 76*(4), 642-667.

Darling-Hammond, L., Hightower, A., Husbands, J. L., LaFors, J. R., Young, V. M., & Christopher, C. (2005). *Instructional leadership for systemic change: The story of San Diego's reform* (Vol. 3). Lanham, MD: Scarecrow Education.

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us. *Educational Researcher, 31*(9), 13-25.

Datnow, A. (2005). The sustainability of comprehensive school reform models in changing district and state contexts. *Educational Administration Quarterly, (41)*1, 121-153.

Datnow, A., Borman, G., Stringfield, S., Rachuba, L., & Castellano, M. (2003). Comprehensive school reform in culturally and linguistically diverse contexts: Implementation and outcomes from a four-year study. *Educational Evaluation and Policy Analysis, 25*(2), 143-170.

Datnow, A., & Stringfield, S. (2000). Working together for reliable school reform. *Journal of Education for Students Placed at Risk,* 5(1), 183-204.

De Wit, D. J., Karioja, K., & Rye, B. J. (2010). Student perceptions of diminished teacher and classmate support following the transition to high school: Are they related to declining attendance? *School Effectiveness and School Improvement, 21*(4), 451-472.

DuFour, R. P., DuFour, R. B., Eaker, R. E., & Karhanek, G. (2004). *Whatever it takes: How professional learning communities respond when kids don't learn.* Bloomington, IN: National Educational Service.

DuFour, R., & Eaker, R. E. (1998). *Professional learning communities at work: Best practices for enhancing student achievement*. Bloomington, IN: National Education Service.

Earl, L., Torrance, N., & Sutherland, S. (2006). Changing secondary schools is hard: Lessons from 10 years of school improvement in the Manitoba school improvement program. In A. Harris & J. Chrispeels (Eds.), *Improving schools and educational systems*. London, GB: Routledge.

European Commission. (2009). *Progress Towards the Lisbon Objectives in Education and Training: Indicators and Benchmarks 2009*. Retrieved from http://ec.europa.eu/education/lifelong-learningpolicy/doc/report09/report_en.pdf

Forte, E. (2010). Examining the assumptions underlying the NCLB federal accountability policy on school improvement. *Educational Psychologist*, *45*(2), 76-88.

Fuchs, D. & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*(1), 93-99.

Fullan, M. (2000). The return of large-scale reform. *Journal of Educational Change, 1*(1), 5-28.

Fullan, M. (2001). *The new meaning of educational change* (3rd ed.). New York, NY: Teachers College Press.

Fullan, M. (2011). *Choosing the wrong drivers for whole system reform* (Report 204). Victoria, Australia: Centre for Strategic Education.

Fullan, M., Bertani, A., & Quinn, J. (2004). New lessons for districtwide reform: Effective lessons for change at the district level has 10 crucial components. *Educational Leadership, 61*(7), 42-46.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation. *Review of Educational Research, 47*(2), 335-397.

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, *95*(1), 148-162.

Gallup (2011). In *New Record-Low Confidence in US Public Schools.* Retrieved from http://www.gallup.com/poll/148724/near-record-low-confidence-public-schools.aspx

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources, 32*(3), 505-523.

Gordon, E. (2009). The global talent crisis. *Futurist*, *43*(5), 34-39.

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching: Theory and Practice, 8*(3/4), 381-391.

Guskey, T. R. (2007). Multiple sources of evidence: An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measures: Issues and Practice, 26*(1), 19-27.

Guskey, T. R. (2009*). Practical solutions for serious problems in standards based grading*. Thousand Oaks, CA: Corwin Press.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2009). New evidence about Brown v. Board of Education: The complex effects of school racial composition on achievement. *Journal of Labor Economics, 27*(3), 349-383.

Haponstall, K. (2010). *An analysis of the correlation between standards-based, non-standards-based grading systems and achievement as measured by the Colorado student assessment program (CSAP).* (Unpublished doctoral dissertation). Capella University, Minneapolis, MN.

Hargreaves, A., & Fink, D. (2006). *Sustainable leadership.* San Francisco, CA: Josey-Bass/Wiley.

Harlow, C. (2003). *Education and correctional populations*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. Retrieved from http://www.policyalmanac.org/crime/archive/education_prisons.pdf

Hattie, J. (2003, October). *Teachers make a difference: What is the research evidence?* Paper presented at the Australian Council for Educational Research Annual Conference, Melbourne, Australia. Retrieved from: http://research.acer.edu.au/ research_conference_2003/4

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* New York, NY: Routledge.

Hess, A. G. (2003). *Changes in student achievement in Illinois and Chicago, 1990-2000.* Chicago, IL: Northwestern University.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin.

Hochbein, C. D. & Duke, D. L., (2011). Crossing the line: Examination of student demographic changes concomitant with declining academic performance in elementary schools. *School Effectiveness and School Improvement, 22*(1), 87-118.

Holdzkom, D. (2002). *Effects of comprehensive school reform in 12 schools: Results of a three-year study.* Appalachian Educational Laboratory: Author.

Honig, M. I., & Hatch, T. C. (2004). Crafting coherence: How schools strategically Manage multiple, external demands. *Educational Researcher, 33*(8), 16-30.

Hopkins, D. & Reynolds, D. (2010). The past, present and future of school improvement: Towards the third age. *British Educational Research Journal, 27*(4), 459-475.

Hoxby, C. (2003). School choice and school productivity: Could school choice be a tide that lifts all boats? In: Hoxby, C. (Eds.), *The economics of school choice* (pp. 287-341). Chicago, IL: University of Chicago Press.

Jefferson County Public Schools (2011a). *Project proficiency guide.* Unpublished manuscript: Author.

Jefferson County Public Schools (2011b). *Student progression, promotion, and grading.* Louisville, KY: Author.

Jencks, C. (1972). *Inequality: A reassessment of the effect of family and schooling in America.* New York, NY: Harper & Row.

Kedro, M. J. (2004). Coherence: When the puzzle is complete. *Principal Leadership, 4*(8), 28-32.

Kemple, J. J., & Willner, C. J. (2008). *Career academies: Long-term impacts on labor market outcomes, educational attainment, and transitions to adulthood*. New York, NY: MDRC.

Kentucky Department of Education. (2008). *Commonwealth accountability testing system: 2007-08 technical report*. Retrieved from http://www.education.ky.gov/kde/administrative+resources/testing+and+reporting+/kentucky+school+testing+system/accountability+system/technical+manual+2008.htm

Kentucky Department of Education. (2010). *No Child Left Behind (NCLB) interpretive guide 2010.* Retrieved from http://www.education.ky.gov/nr/rdonlyres/0a2e4cd2-7b79-476c-a16a-33415da5e2fe/0/2010_nclb_interpretive_guide.pdf

Kentucky Department of Education. (2011). *Kentucky Department of Education Open House.* Retrieved from http://openhouse.education.ky.gov/

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling.* London, GB: Sage Publications.

Land, D., & Legters, N. (2002). The extent and consequences of risk in U. S. education. *Yearbook of the National Society for the Study of Education, 101*(2), 1-288.

Lasky, S., Stringfield, S., Teddlie, C., Kennedy, E., Schaffer, E., Chrispeels, J., Daly, A., McDonald, D. (2005). Designing and conducting a gold standard effective schools study. *Journal for Effective Schools, 4*(1), 27-45.

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62*(3), 172-192.

Lekholm, A. K., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school levels: Effects of gender and family background. *Educational Research and Evaluation*, *14*(2), 181-199.

Levin, H. M. (2001). *Privatizing education: Can the marketplace deliver choice, efficiency, equity, and social cohesion?* Boulder, CO: Westview Press.

Levin, H., Belfield, C., Muennig, P., & Rouse, C. (2007). *The costs and benefits of an excellent education for all of America's children.* New York, NY: Columbia University, Teachers College.

Linn, R.L. (2009). The concept of validity in the context of NCLB. In Lissitz, R. W., *The concept of validity: Revisions, new directions, and applications* (195-212). Charlotte, NC: Information Age Publishing.

Lubienski, C. (2003). Innovation in education markets: Theory and evidence on the impact of competition and choice in charter schools. *American Educational Research Journal, 40*(2), 395-443.

Lybbert, B. (1998). *Transforming learning with block scheduling: A guide for principals*. Thousand Oaks, CA: Corwin Press.

Lynch, M., & Cicchetti, D. (1997). Children's relationships with adults and peers: An examination of elementary and junior high school students. *Journal of School Psychology, 35*(1), 81-99.

MacIver, M. A. (2010). *Gradual disengagement: A portrait of the 2008-09 dropouts in the Baltimore city schools*. Baltimore, MD. Baltimore Education Research Consortium. Retrieved from http://www.acy.org/upimages/Gradual_ Disengagement.pdf

MacIver, M. A. (2011). The challenge of improving urban high school graduation outcomes: Findings from a randomized study of dropout prevention efforts. *Journal of Education for Students Placed at Risk, 16*(3), 167-184.

MacIver, M. A., Durham, R. E., Plank, S. B., Farley-Ripple, E. F., & Balfanz, R. (2007). *The challenge of on-time arrival: The seven-year flight paths of Baltimore's sixth graders of 1999-2000*. Baltimore, MD: Baltimore Education Research Consortium. Retrieved from http://www.every1graduates.org/PDFs/SIXTH%20pathways.pdf

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal, 37*(1), 153-184.

Marsh, H, & Craven, R. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*(2): 133-163.

Marsh, H., & O'Mara, A. (2008). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept, personality, and social psychology. *Personality and Social Psychology Bulletin, (34)*4, 542-552.

Marzano, R. J. (2000). *Transforming classroom grading*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2010). *Formative assessment & standards-based grading*. Bloomington, IN: Marzano Research Laboratory.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for every teacher.* Alexandria, VA: ASCD.

McDonald, J. P., Mohr, N., Dichter, A., & McDonald, E. C. (2007). *The power of protocols: An educator's guide to better practice* (2nd ed.). New York, NY: Teacher's College Press.

McLaughlin, M. W. (1990). The RAND change agent study revisited: Macro perspectives and micro realities. *Educational Researcher, 19*(9), 11-16.

McLaughlin, M., & Mitra, D. (2001). Theory-based change and change-based theory: going deeper, going broader. *Journal of Educational Change*, *2*(4), 301-323.

McMillan, J. H. (2001). Secondary teacher's classroom and grading practices. *Educational Measurement: Issues and Practices, 20*(1), 20-32.

McMillan, J. H., & Nash, S. (2000, April). *Teacher classroom assessment and grading practice and decision making*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

McPartland, J., Balfanz, R., Jordan, W., & Legters, N. (1998). Improving climate and achievement in a troubled urban high school through the talent development model. *Journal of Education for Students Placed At Risk, 3*(4), 337-361.

Mickelson, R. A. (2010). Goals, grades, fears, and peers. Introductory essay for special issues on the effects of school and classroom racial and SES composition on educational outcomes. *Teachers College Record, 112*(4), 961-977.

Minimum Requirements for High School Graduation, 704 KAR 3:305 (2011). Retrieved from http://www.lrc.ky.gov/kar/704/003/305.htm

Mishook, J., Foley, E., Thompson, J., & Kubiak, M. (2008). Beyond test scores: Leading indicators for education. *Voices in Urban Education* (Winter 2008). Retrieved from http://www.annenberginstitute.org/VUE/winter08/Mishook.php

Muñoz, M. A., & Chang, F. C. (2007). The elusive relationship between teacher characteristics and student academic growth: A longitudinal multilevel model for change. *Journal for Personnel Evaluation in Education, 20*(1), 147-164.

Muñoz, M. A., Guskey, T. R., & Aberli, J. R. (2009). Struggling readers in urban high schools: Evaluating the impact of professional development in literacy. *Planning and Changing, 40*(1/2), 61-85.

Nathan, J. (1999). *Charter schools: Creating hope and opportunity for American education*. San Francisco, CA: Jossey-Bass.

National Assessment of Education Progress. (2011). Retrieved from http://nces.ed.gov/nationsreportcard/pdf/about/2011471.pdf

National Center on Education and the Economy (2008). *Tough choices or tough times: The report on the new commission on the skills of the American workforce, revised and expanded.* San Francisco, CA: Jossey-Bass.

Neild, R. C., & Balfanz, R. (2006). *Unfulfilled promise: The dimensions and characteristics of Philadelphia's dropout crisis, 2000-2005*. Philadelphia Youth Network, Johns Hopkins University, and University of Pennsylvania.

Neild, R. C., Balfanz, R., & Herzog, L. (2007). An early warning system. *Educational Leadership, 65*(2), 28-33.

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis, 23*(4), 297-321.

No Child Left Behind Act, 20 U.S.C. § 6319 (2001).

Nunnery, J. A. (1998). Reform ideology and the locus of development problem in educational restructuring. *Education and Urban Society, 30*(3), 277-295.

Odden, A., Borman, G., & Fermanich, M. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education, 79*(4), 4-32.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

Oxley, D. (2008). Creating instructional program coherence. *Principal's Research Review, 3*(5), 1-7.

Oxley, D., & Kassissieh, J. (2008). From comprehensive high schools to small learning communities: Accomplishments and challenges. *Forum, (50)*2, 199-206.

Oxley, D., & Luers, K. W. (2011). How small schools grew up and got serious (but did not lose their spunk). *Phi Delta Kappan, 92*(4), 62-66.

Palardy, G. J. (2008). Differential school effects among low, middle, and high social class composition schools: A multiple group, multilevel latent growth curve analysis. *School Effectiveness and School Improvement, 19*(1), 21-49.

Payne, C. M. (2008). *So much reform, so little change: The persistence of failure in urban schools*. Cambridge, MA: Harvard Education Press.

Perry, L. B., & McConney, A. (2010). Does SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record, 112*(4), 1137-1162.

Persistently Low-Achieving School and School Intervention Defined. Kentucky Revised Statutes 160.346 (2010).

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management, 26*(4), 909-949.

Popham, J. (1997). The moth and the flame: Student learning as a criterion of instructional competence. In J. Millman (Ed.), *Grading teachers, grading schools. Is student achievement a valid education measure?* (pp. 264-264). Thousand Oaks, CA: Corwin.

Quint, J. (2006). *Meeting five challenges of high school reform: Lessons from research on three reform models*. Manpower Demonstration Research Corporation: Author.

Race to the Top Act, 75. Fed Reg. 19496 (2010).

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* Princeton, NJ: Educational Testing Service.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2011). HLM 7 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Sciences, 20*(4), 307-355.

Ravitch, D. (2011, November). NCLB: End it, don't mend it. *Education Week,* p. 9.

Ravitch, D. (2010). *The death and life of the great American school system*. New York, NY: Basic Books.

Reschly, A. L., Huebner, E. S., Appleton, J. J., & Antaramian, S. (2008). Engagement as flourishing: The contribution of positive emotions and coping to adolescents' engagement at school and with learning. *Psychology in the Schools, 45*(5), 419-431.

Reynolds, D., Creemers, B., Stringfield, S., Teddlie, C., & Schaffer, G. (Eds.). (2002). *World class schools: International perspectives on school effectiveness*. London, GB: Routledge Falmer.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rorrer, A. K., Skrla, L., & Schuerich, J. J. (2008). Districts as institutional actors in educational reform. *Educational Research Quarterly, 44*(3), 307-358.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Corwin.

Rumberger, R. W., & Lim, S. (2008). *Why students drop out of school: A review of 25 years of research*. (California Dropout Project Research Report #15) Santa Barbara, CA: California Dropout Research Project. Retrieved from http://www.youblisher.com/p/135474-Why-Students-Drop-Out-of-School-A-Review-of-25-Years-of-Research/

Rumberger, R. W., & Pallardy, G. J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record, 107*(9), 1999-2045.

Sanders, M. G. (2012). Achieving scale at the district level: A longitudinal multiple case study of a partnership reform. *Educational Administration Quarterly, 48*(1), 154-186.

School Improvement Fund (2010). *School Improvement Grants; American Recovery and Reinvestment Act of 2009 (ARRA); Title I of the Elementary and Secondary Education Act of 1965, as Amended (ESEA)***.**Retrieved from http://www2.ed.gov/programs/sif/legislation.html

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Silver, D., Saunders, M., & Zarate, E. (2008). *What factors predict high school graduation in the Los Angeles unified school district?* California Dropout Research Project Report #14. UCLA/IDEA and UC/ACCORD. Retrieved from http://www.edcoe.org/departments/curriculum_instruction/documents/CILC082008_DropOutFactors.pdf

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and timing of events. *Journal of Educational Statistics, 18*(2), 155-195.

Slavin, R. E., Madden, N. A., Dolan, L. J., Wasik, B. A., Ross, S., Smith, L., & Dianda, M. (1996). Success for all: A summary of research. *Journal of Education for Students Placed at Risk, 1*(4), 41-76.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233-267). Bristol, PA: Falmer.

Stiggins, R. J. (2007). Assessment through the student's eyes. *Educational Leadership, 64*(8), 22-26.

Stiggins, R. J. (2008). *Assessment Manifesto: A Call for the Development of Balanced Assessment Systems.* Princeton, NJ: Educational Testing Service.

Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning: Doing it right—using it well.* Portland, OR: Assessment Training Institute.

Stiggins, R. J., & DuFour, R. (2009). Maximizing the power of formative assessments. *Phi Delta Kappan, 90*(9), 640-644.

Stiggins, R. J., Frisbie, D. A., & Griswold, P.A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8*(2), 5-14.

Stringfield, S., & Datnow, A. (1998). Scaling up school restructuring designs in urban schools. *Education and Urban Society, 30*(3), 269-276.

Stringfield, S., & Datnow, A. (2002). Systemic supports for schools serving students placed at risk. In S. Stringfield & D. Land (Eds.), *Educating at-risk students* (pp. 269-285). Chicago, IL: National Society for the Study of Education.

Stringfield, S., Datnow, A., Ross, A., & Snively, F. (1998). Scaling up school restructuring in multicultural, multilingual contexts: Early observations from Sunland County. *Education and Urban Society, 30(*3*),* 326-357.

Stringfield, S., Reynolds, D., & Schaffer, E. C. (2008). Improving secondary students' academic achievement through a focus on reform reliability: 4- and 9-year findings from the high reliability schools project. *School Effectiveness and School Improvement, 19*(4), 409-428.

Stringfield, S., Reynolds, D., & Schaffer, E. C. (2010, October). *Toward highly reliable, high-quality public schooling*. Paper presented at the McREL Best in the World Consortium Meeting, Denver, CO.

Stringfield, S., & Yakimowski-Srebnick, M. E. (2005). Promise, progress, problems, and paradoxes of three phases of accountability: A longitudinal case study of the Baltimore City public schools. *American Educational Research Journal, 42*(1), 43-75.

Stronge, J. H. (2010). *Effective teachers = student achievement: What the research says.* Larchmont, NY: Eye on Education.

Supovitz, J. A. & Weinbaum, E. H. (Eds.). (2008). *The Implementation Gap: Understanding reform in high schools.* New York, NY: Teachers College Press.

Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Teddlie, C., Stringfield, S., Wimpleberg, R., & Kirby, P. (1987, April). *Contextual differences in effective schooling in Louisiana*. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC. Retrieved from ERIC database. (ED 286272).

Tewke, C. D., Carter, R. L., Chang-Xing, M., Algina, J., Lucas, M. E., Roth, J., . . . Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11-35.

Toch, T. (2003). *High schools on a human scale: How small schools can transform American education.* Boston, MA: Beacon Press.

Tomlinson, C. A., & McTighe, J. (2006). *Integrating differentiated instruction & understanding by design: Connecting content and kids*. Alexandria, VA: Association for Supervision and Curriculum Development.

Trochim, W. M., & Donnelly, J. P. (2008). *The research methods knowledge base* (3rd ed.). Mason, OH: Atomic Dog Publishing.

Tucker, M. (2011). *Standing on the shoulders of giants: An American agenda for educational reform.* Washington, DC: National Center on Education and the Economy.

United States Department of Education, National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. A report to the Nation and the Secretary of Education. Washington, D.C.

Useem, E., Offenberg, R., & Farley, E. (2007). *Closing the teacher quality gap in Philadelphia: New hope and old hurdles*. A report from Learning from Philadelphia's School Reform. Philadelphia, PA: Research in Action.

Wald, J., & Losen, D. (Eds.) (2003). *Deconstructing the school to prison pipeline.* New Directions for Youth Development (Number 99). San Francisco, CA: Jossey-Bass.

Wentzel, K. R. (1997). Student motivation in middle school: The role of perceived pedagogical caring. *Journal of Educational Psychology, 89*(3)*,* 411–419.

Whitlock, J. L. (2004). *Places to be and places to belong: Youth connectedness in school and community.* Retrieved from http://ecommons.library.cornell.edu/bitstream/1813/19327/2/places.pdf

Wilkins, J. (2008). School characteristics that influence student attendance: Experiences of students in a school avoidance program. *The High School Journal, 91*(3), 12-24.

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record, 112*(4), 1008-1037.

Winn, D., Menlove, R., & Zsiray, S. W. (1997). An invitation to innovation: Rethinking the high school day. *NASSP Bulletin, 81*(588), 10-18.

Xiang, Y., & Hauser, C. (2010). *School conditional growth models: How to make an "apples to apples" comparison possible?* Northwest Evaluation Association: Author.

Zavadsky, H. (2009). *Bringing school reform to scale: Five award-winning urban districts*. Cambridge, MA: Harvard Education Press.

Zvoch, K. (2006). Freshman year dropouts: Interactions between student and school characteristics and student dropout status. *Journal of Education for Students Placed at Risk, 11*(1), 97–117.

Appendix 1

HLM Models: Fixed and Random Effects and Model Fit

	Mathematics (Project Proficiency)			Social Studies (Non-Project Proficiency)		
Fixed Effects	*Coefficient*	*se*	*t Ratio*	*Coefficient*	*se*	*t Ratio*
Unconditional Model						
Average classroom mean, γ_{00}	28.50	.73	38.94***	27.49	.60	45.50***
Means-as-Outcomes Model						
INTERCEPT, γ_{00}	24.01	.87	27.60***	26.64	.91	29.27***
ClassSES, γ_{01}	-13.30	4.02	-3.31**	-12.84	3.62	-3.55***
PP, γ_{02}	8.58	6.76	6.80***	1.57	1.17	1.34
Random Coefficient Model						
Overall mean achievement, γ_{00}	12.08	.70	17.17***	10.64	.80	13.16***
Mean StudentSES achievement slope, γ_{10}	-.54	.50	-1.08	-.90	.56	-1.60
Mean Prior Achievement slope, γ_{20}	.60	.01	43.01***	.58	.02	32.04***
Intercepts- and Slopes-as-Outcomes Model						
Model for classroom means						
INTERCEPT, γ_{00}	7.59	.94	8.05***	8.43	1.20	7.03***
ClassSES, γ_{01}	.54	3.68	.15	-1.12	5.25	-.21
PP, γ_{02}	8.00	1.32	6.04***	3.32	1.58	2.11*
Model for StudentSES-achievement slopes						
INTERCEPT, γ_{10}	-.59	.75	-.78	-2.00	.83	-2.39*
ClassSES, γ_{11}	-3.04	3.01	-1.01	-.50	3.78	-.13
PP, γ_{12}	.20	.99	.20	2.64	1.11	2.40*

Model for Prior Achievement slopes								
INTERCEPT, γ_{20}		.60	.02	30.57***		.62	.02	26.53***
ClassSES, γ_{21}		-.16	.08	-2.17*		-.10	.11	-.99
PP, γ_{22}		.03	.03	1.21		.09	.03	-2.47*
Random Effects	*Variance*	*df*	X^2	*p Value*	*Variance*	*df*	X^2	*p Value*
Unconditional Model								
Classroom mean, u_{0j}	90.02	211	1169.02	***	58.10	211	971.11	***
Level-1 effect, r_{ij}	203.37				170.97			
Means-as-Outcomes Model								
Classroom mean, u_{0j}	61.38	209	833.84	***	51.50	209	864.50	***
Level-1 effect, r_{ij}	204.80				171.37			
Random-Coefficient Model								
Classroom mean, u_{0j}	32.23	176	231.33	*	31.73	178	216.36	*
StudentSES achievement slope, u_{1j}	.45	176	175.38	.50	8.65	178	180.49	.43
Prior achievement slope, u_{2j}	.00	176	164.70	.50	.01	178	206.93	.07
Level-1 effect, r_{ij}	122.21				116.40			
Intercepts- and Slopes-as-Outcomes Model								
Classroom mean, u_{0j}	20.39	174	207.94	**	31.17	176	215.34	*
StudentSES slope, u_{1j}	.50	174	173.89	.50	6.90	176	173.20	.50
Prior achievement slope, u_{2j}	.00	174	165.00	.50	.01	176	205.94	.06
Level-1 effect, r_{ij}	123.33				116.35			

Model Fit	*Deviance*	*Parameters*	*Deviance*	*Parameters*
Unconditional Model	20340.92	2	19872.98	2
Means-as-Outcomes Model	20286.93	2	19851.18	2
Random-Coefficient Model	19045.13	7	18856.53	7
Intercepts- and Slopes-as-Outcomes	18924.94	7	18815.15	7

Note. *p < .05. **p< .01. ***p< .001.

CURRICULUM VITAE

NAME: Joseph C. Burks, Jr.

ADDRESS: 9400 Exhibition Court
Louisville, KY 40291

DOB: Hamilton AFB, California- August 4, 1953

EDUCATION
& TRAINING: B.S., Mathematics
Centre College
1971-1975

M.Ed., Secondary Education
University of Louisville
1975-1978

AWARDS: 1994 Kentucky Department of Education State Principal of the Year

2011 Gheens Institute for Innovation Award

SELECTED PRESENTATIONS:

2007 Kentucky Association for Assessment Coordinators: Learning from Student Work

2008 America's Promise Dropout Summit – Bellarmine College, Louisville, KY: Reinventing the High School Structure

2010 AdvancEd Innovation Summit – Kentucky P20 Innovation Lab: Project Proficiency and Guaranteed Student Competency

2011 KentuckianaWorks Board: JCPS Professional Career Themes

2011 55,000 Degrees Board: JCPS implementation of National Common Core Standards

CURRICULUM VITAE

NAME: Glenn Stephen Baete

ADDRESS: 7905 Ridgehurst Place
Louisville, KY 40299

DOB: Louisville, Kentucky- December 26, 1968

EDUCATION
& TRAINING: B.A., English
University of Louisville
1986-1991

M.Ed., Secondary Education
University of Louisville
1991-1996

AWARDS: Outstanding High School Principal of the Year, Jefferson County Public Schools 2008

PUBLICATIONS: Baete, G. S. (2011). Book Review [Review of the Book *Differentiating school leadership: Facing the challenges of practice, by D. L. Duke]*. *School Leadership and Management, 31*(5), 545-548.

NATIONAL MEETING PRESENTATIONS:

United States Department of Education (USDOE) - Smaller Learning Communities Project Director Meeting, November, 2011
Topic: Professional Collaboration

Education Northwest: "From Structures to Instruction" National Conference, June 2011
Topic: District Supports to Improve Student Learning

INVITED PRESENTATIONS:

Great Schools Partnership/USDOE- Smaller Learning Communities Webinar, March 2011
Topic: Harnessing Teacher Knowledge through Professional Collaboration

CURRICULUM VITAE

NAME: Martin Anthony Pollio

ADDRESS: 2352 Page Avenue
Louisville, KY 40205

DOB: Louisville, Kentucky- June 21, 1971

EDUCATION
& TRAINING: B.S., Education
Indiana University
1989-1993

M.Ed., Secondary Education
Eastern Kentucky University
1993-1995

AWARDS: Russell Garth Award Winner, Outstanding Principal, Jefferson County Public Schools 2011

NATIONAL MEETING PRESENTATIONS:

Education Northwest: "From Structures to Instruction" National Conference, June 2011
Topic: District Supports to Improve Student Learning

High Schools that Work: National Conference, July 2011
Topic: Developing a Culture of Instructional Improvement

Kentucky Association of School Administrators: Showcase of Kentucky Schools, March 2012
Topic: Why Do You Have a Successful School?