Evaluating Teacher Effectiveness in a Professional Development Program: Considering Measures for Inclusion in a Comprehensive Teacher Evaluation System

MacCalla, Nicole Marie-Gerardi

2014

UNIVERSITY OF CALIFORNIA

Los Angeles

Evaluating Teacher Effectiveness in a Professional Development Program:

Considering Measures for Inclusion in a Comprehensive Teacher Evaluation System

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Education

by

Nicole Marie-Gerardi MacCalla

2014

ASBRACT OF THE DISSERTATION

Evaluating Teacher Effectiveness:

Investigating Measures for Inclusion in a Comprehensive Teacher Evaluation System

by

Nicole Marie-Gerardi MacCalla

Doctor of Philosophy in Education

University of California, Los Angeles, 2014

Professor Marvin C. Alkin, Chair

Effective teachers not only affect academic achievement but also a lifetime of success. Equal opportunity to access quality education is recognized as a fundamental constitutional right for every child in America, yet it is rarely actualized. Supporting the belief that teaching matters, school reform measures turn to focusing on ensuring access to effective teaching. The federal government, states, and districts, are proposing new teacher evaluation systems with inherently higher levels of validity and reliability that can more accurately and meaningfully assess teacher effectiveness.

Comprehensive teacher evaluation systems (CTESs) use multiple-measures of teacher effectiveness (e.g. student achievement, personnel review, self-and-peer-evaluation, student feedback, etc.) that differentiate levels of teaching, for formative (improve teaching and learning) and summative (decision-making) purposes. As longstanding perfunctory evaluation systems are replaced by CTESs, discussions revolve around the “right” measures for inclusion.

Using pre-existing data (teacher survey, expert assessment, classroom observation), from a three-year state funded Improving Teacher Quality (ITQ) science and social studies-history, urban middle school, professional development (PD) program, this study explores: the sensitivity of measures to detecting differences (within-groups and between-groups); relationships between teacher effectiveness constructs (and teacher characteristics and PD); and the extent to which depictions of teachers vary across different measures of effectiveness. This study takes a step in understanding what measures should be included in a CTES aimed a providing a complete assessment of teacher quality.

Findings indicate low-to-moderate-levels of sensitivity in detecting differences and high-levels of construct score consistency within the expert assessment and classroom observation measures. Further, the validity and reliability of the teacher survey is questioned, eliminating it for consideration in a CTES. Teacher characteristics do a poor job predicting scores on teacher effectiveness constructs, while PD participation and use of instructional strategies moderately predict construct scores. Classroom observations provide a unique portrayal of teacher effectiveness and are strongly recommended for inclusion in a CTES.

While comprehensive teacher evaluation systems take a holistic approach to evaluation, recognizing and valuing the complexities of teaching, connecting teacher performance to personnel decisions revolutionizes education. As we move towards identifying, retaining, rewarding, and developing effective teachers, the promise of public education is hopeful.

The dissertation of Nicole Marie-Gerardi MacCalla is approved.

Christina A. Christie

Todd M. Franke

Tyrone C. Howard

Marvin C. Alkin, Committee Chair

University of California, Los Angeles

2014

DEDICATION

To all the friends and family that have helped make this dream a possibility. I could not have done this without your love, support, and unwavering belief in me. This is our victory!

I am especially grateful:

To my Mom Val, for planting the seeds of education and helping them grow.
To Gerald Brunetti, for believing in me and encouraging me to go to graduate school.
To my Aunt Re, for her stability, support, and adventures. From the cradle to the grave!
To Mama Christine, my soul sister and support system, for keeping me sanely progressing.
To Esther, My Angel, for loving my children and keeping my sanctuary pristine!
To Jackie and Thomas, for keeping my family thriving during that final push.
To my Husband Reuben, for clearing the way and keeping me well nourished and supported.
To my Son Sebastian, for making me laugh and hanging in there with me every step of the way!
To my Daughter Ande Jaye, for keeping me sweet and smiling.
To source, for which all things come.

This one is for us!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

Acronym or Abbreviation	Full Name	Description
APD	Academic Year Professional Development	A combination of both forms of professional development (ADPD and EDPD) offered to teachers during the school year as part of the ITQ Program.
ADPD	All-Day Professional Development	A 7 hour form of professional development offered to teachers on weekends approximately 3 to 4 times per year as part of the ITQ Program.
API	Academic Performance Index	An academic performance characteristic used to describe the schools in the study.
AYP	Adequate Yearly Progress	An academic performance characteristic used to describe the schools in the study.
CDE	California Department of Education	The final ITQP funding agency (program and evaluation) and a source for publically available data on schools.
CFA	Confirmatory Factor Analysis	Statistical analysis used to test whether the Teacher Survey data fits the hypothesized measurement model.
COP	Classroom Observation Protocol	An outcome measure used in the study to assess teaching practice.
CPEC	California Postsecondary Education Commission	The original funding agency for the ITQ Program and evaluation.
CST	California Standards Test	An outcome measure originally envisioned for use in this study to assess 8th grade mastery of science and social studies standards.
CTES	Comprehensive Teacher Evaluation System	A multi-measure teacher evaluation system offering a holistic view of teacher quality through valid and reliable estimates of teacher effectiveness.
EA	Expert Assessment	An outcome measure used in this study to assess teaching practice.
EDPD	Extended-Day Professional Development	A 1.5 hour form of professional development offered to teachers after school roughly 14 times during the school year as part of the ITQ program.
ELL	English Language Learner	A characteristic used to describe students at the schools in the ITQP study.
FRPL	Free and Reduced-Price Lunch	A measure of socio-economic status used for students at the schools in the ITQP study.
ITQ	Improving Teacher Quality	The name of the state grants program responsible for funding the ITQ Program.
ITQP	Improving Teacher Quality Program	The generic name for the program from which data is obtained for the study.
LA	Los Angeles	The city in which ITQP services took place.

Acronym or Abbreviation	**Full Name**	**Description**
LASD	Los Angeles Unified School District	The school district where ITQP services took place.
M	Mean	Average.
PD	Professional Development	A service to teachers in the ITQ Program.
PLAS	Partnership for Los Angeles Schools	A subset of schools sponsored by and targeted for improvement by the Mayor of Los Angeles where program services took place.
PLP	Professional Learning Partner	The day-to-day provider of ITQP services.
PI Status	Program Improvement Status	An academic performance characteristic used to describe the schools in the study.
RFEP	Reclassified Fluent-English Proficient	A characteristic used to describe student demographics at the schools in the ITQP study.
RTI	Response to Intervention	A model used in the delivery of the ITQP professional development, which assesses student learning to drive instruction and professional development.
R2T	Race to the Top	A federal funding stream with explicit guidelines for teacher evaluation.
SD	Standard Deviation	A statistic that describes variation around the mean.
SPD	Summer PD	A form of ITQP services that occurred over the summer.
SWD(s)	Students with Disabilities	A characteristic used to describe students at the schools in the ITQP study.
T	Time	Referring to the data collection time-point for each instrument.
T&L Framework	Teaching and Learning Framework	A modified version of Charlotte Danielson's Framework for Teaching used in the LAUSD teacher evaluation system.
TS	Teacher Survey	An outcome measure used in the study to assess teaching practice.
UCLA	University of California, Los Angeles	The Institution of Higher Learning partnered with PLAS and LAUSD that received funds from CPEC/CDE to administer the ITQ Program.
UCLA SRM Evaluation Group	University of California, Los Angeles - Social Research Methodology Evaluation Group	The party responsible for conducting the evaluation study of the ITQ Program.

ACKNOWLEDGEMENTS

I would like to thank UCLA Center X Directors, Staff, and Professional Learning Partners for help in designing instruments and facilitating data collection for the Improving Teacher Quality (ITQ) Program and subsequently making data available for use in this study. Your dedication to improving education is what makes this study possible.

Thank you to the California Department of Education (CDE) and the California Postsecondary Education Commission (CPEC) for funding the ITQ Program and evaluation study and subsequently making the data available for use in this study. Your vision for education is what makes this study possible.

I would also like to thank all of the teachers that participated in the ITQ Program and subsequent studies. Your trust and candidness is what makes this study successful.

I send my gratitude to UCLA SRM Evaluation Group for giving me a home to develop and practice my craft for so long! I have learned so much from my tenure with the group and I will be forever thankful for all that comes my way because of the proficiencies that I have developed with their support. Thank you to all of the members, past and present, for your mentorship and for your help in conceptualizing workshops, studies, instruments, and analyses. I deeply value your support and expertise. I will miss you all.

I extend my deepest gratitude to my dissertation committee for their invaluable insights and guidance in conceptualizing and conducting this study. I know it has been a long arduous process and I appreciate the space that has been granted to me to finish the dissertation. Thank you for helping me to think carefully about the study design and analysis. Your feedback has

greatly enhanced the quality of this investigation. A special thank you goes to Jody Priselac, a non-signing committee member, for your unwavering understanding, support, and advocacy.

A special thank you goes out to Marv, Director of the SRM Evaluation Group, Professor Emeritus, Advisor, and Committee Chair. You have helped me believe in myself and shown me what I am capable of. Thank you for all that you have taught me and for all of the opportunities that you have given me. Your unyielding guidance and support is what has gotten me this far. Thank you from the bottom of my heart!

I have the utmost respect and appreciation for Lisa Dillman, SRM Evaluation Group colleague, ITQ Co-Director, Co-Editor, Co-Author, Co-Presenter, and Friend. You picked up where I left off and helped pull me over the finish line (more than once)! Thank you for sharing your depth of expertise with me, for motivating me, and for illuminating the humor in all of our co-adventures. You will always have my highest recommendation. I wish you the absolute best in all things!

Lastly, I acknowledge the funding and support received from UCLA's Graduate Division, UCLA Graduate School of Education & Information Studies (GSE&IS) and the Social Research Methodology Division (SRM) within GSE&IS. I am indebted to you for your continuing guidance and support. Thank you for giving me the space to develop into the valuable professional that I am today.

VITA

EDUCATION

2002	B.A. Economics and Liberal & Civic Studies Saint Mary’s College of California
2006	M.A. Education University of California, Los Angeles (UCLA)

SELECT HONORS AND AWARDS

2002	Adam Smith Award Alfred Fromm Award Saint Mary’s College of California
2007-2014	UC Regents University Fellowship University of California, Los Angeles

PROFESSIONAL EXPERIENCE

2006-2014	Graduate Student Researcher UCLA (Social Research Methodology) SRM Evaluation Group
2007-2009	Teaching Evaluation Section Co-Editor *American Journal of Evaluation*
2008	Professional Expert Los Angeles Unified School District (LAUSD)
2008-2014	Evaluation Capacity Building Workshop Facilitator UCLA Academic Preparation and Educational Partnerships (APEP)
2009-2014	Director of Research Improving Teacher Quality Programs with UCLA Center X
2010-2011	Teaching Assistant/Special Reader UCLA GSE&IS Courses in Evaluation

PROFESSIONAL CERTIFICATION

2010	Cognitive Coach Center for Cognitive Coaching
2011	Classroom Observer Los Angeles Unified School District (LAUSD)

SELECT PRESENTATIONS

Azzam, T., & Gerardi, N. (2006). Evaluating a University and District Partnership. Paper presented at the American Educational Research Association (AERA) Annual Conference – Chicago, Illinois.

Gerardi, Nicole. (2008). Increasing the Utilization of Evaluation Findings through the Disaggregation of Survey Data. Paper presented at the American Evaluation Association (AEA) Annual Conference – Denver, Colorado.

Lee, J., & Gerardi, N. (2011). Practical Considerations for Implementing Longitudinal School-based Evaluations. Paper presented at the American Evaluation Association (AEA) Annual Conference – Anaheim, California.

Dillman, L. M., & Gerardi, N. (2012). Using Technology to Facilitate Participatory Evaluation: A Case Study of a Teacher-Initiated Inquiry Project Evaluation. Paper presented at the American Evaluation Association (AEA) Annual Conference – Minneapolis, Minnesota.

Gerardi, N., & Dillman, L. M. (2012). The Use of a Teacher Survey to Evaluate Increases in Teacher Effectiveness in a Multi-Measure Evaluation System. Poster presented at the American Evaluation Association (AEA) Annual Conference – Minneapolis, Minnesota.

SELECT REPORTS and PUBLICATIONS

Gerardi, Nicole. (2009). Center X / Local District 7 Partnership: Impact Evaluation: Literacy Coach Teacher Survey Results (2007-2008). UCLA SRM Evaluation Group.

MacCalla, N., Dillman, L., & Alkin, M. (2013). Final Evaluation Report: A Quasi-Experimental Design Study of a Professional Development Program for Middle School Science and Social Studies Teachers. UCLA SRM Evaluation Group for the California Department of Education (CDE).

Hipolito, E., & MacCalla, N. (Eds.). (2014). Teacher Collaborative Inquiry. *XChange*. http://centerx.gseis.ucla.edu/xchange-repository

CHAPTER 1: Introduction

"Every child in America deserves a world-class education…. This effort will require the skills and talents of many, but especially our nation's teachers, principals, and other school leaders. Our goal must be to have a great teacher in every classroom and a great principal in every school. We know that from the moment students enter a school, the most important factor in their success is not the color of their skin or the income of their parents – it is the teacher standing in front of the classroom. To ensure the success of our children, we must do better to recruit, develop, support, retain, and reward outstanding teachers in America's classrooms."

(President Barack Obama, March 2010, in *A Blueprint for Reform – The Reauthorization of the Elementary and Secondary Education Act*, p. 1)

In today's educational environment, there is an insatiable interest in documenting and assessing student achievement. Research consistently shows that teacher quality has a direct impact on student achievement well establishing that teaching matters (Darling-Hammond, 2000 and 2011; Ferguson, 1991; Haycock, 1998; Hanuschek et al., 1992; Kemp & Hall, 1992; Rice, 2003; Rivers & Saunders, 1996). The cumulative effect of teacher quality not only affects academic performance, but also a lifetime of success (Chetty, Friedman & Rockoff, 2011; Sanders & Rivers, 1996). Because we know that teaching matters, the evaluation of teacher effectiveness is an important and pressing topic in today's educational system.

Over the past two decades, much of the debate in education has revolved around raising student achievement and improving teacher quality. A succession of reforms has swept across the nation from curricular changes and smaller schools and class sizes to federal requirements of teacher quality and the overhaul of the teacher evaluation system. Longstanding perfunctory evaluation systems are quickly being replaced by comprehensive teacher evaluation systems that recognize the complexities of the teaching profession and attempt to measure teacher effectiveness in valid and reliable ways. Although there is much consensus that teaching matters, there is yet to be consensus on the best way(s) to evaluate teacher performance (Gordon et al., 2006).

In California, evaluation of teacher effectiveness has become a civil rights issue. In the case of Vegara v. California, 9 public school children, joined with Students Matter, claim that the persistence of ineffective teachers, largely in schools serving disadvantaged students, denies every student's basic right and access to quality education. Specifically, the case targets 5 statues of the California Education Code regarding the tenure, dismissal, and layoff processes that do not take into account teacher performance, in effect "creating an unjustifiable and unconstitutional inequality among students" (Students Matter Overview, p. 2). By challenging the systems and processes in place regarding hiring and firing of teachers that inhibit doing what is best for children, the case "seeks to ensure equal access to quality education for all students, embrace a teacher career system that elevates teacher quality, and raises the prestige of the teaching profession" (Students Matter Overview, p.1). While the two-month trial closed on April 10, 2014, a final ruling by the Superior Court judge has not been issued (at the filing of this dissertation). If the case is won, a multi-measure evaluation of teacher effectiveness will be a mandatory component in the hiring, tenure, and dismissal processes for teachers in California.

Statement of the Problem

Traditional teacher evaluation methods include: assessment of teacher qualifications (certification and licensure), and perfunctory personnel evaluation based on short and infrequent observation, which commonly results in a "satisfactory" or "unsatisfactory" rating (Weisberg et al., 2009). These traditional methods offer no specific ways for teachers to improve and have limited, at best ties to hiring, tenure, pay, professional development opportunities, or dismissal

(Weisberg et al., 2009). More recently, researchers, educational institutes, teacher unions, etc. are proposing more comprehensive and holistic ways to measure teacher effectiveness. The suggested measures for inclusion in these models span from student achievement measures (most often value-added estimates, when available); personnel review (most often done by a Principal or other expert); self-evaluation; peer-evaluation; student work; and student feedback. Portfolio Evaluation is believed to be a system with inherently higher levels of validity because it takes a holistic approach, recognizing and valuing the complexities of teaching, in teacher evaluation (Goe, Bell, & Little, 2008).

I am in agreement with Laura Goe, that, "The ultimate goal of teacher evaluation should be to improve teaching and learning" (2010, p. 4). Much of the recent focus (with Value-Added Modeling) has been on evaluating an individual teacher's effect on student achievement. Understanding how a teacher contributes to gains in student achievement is important, however, it is too far removed from the classroom to improve teaching or learning (Goe, 2010). Focusing on measuring gains in student achievement is just one element in understanding teacher effectiveness. Because teaching and learning is a highly complex endeavor, student achievement taken alone, is an inadequate assessment of teacher effectiveness. The urgency, felt in the past few decades, to increase student achievement and improve teacher quality, simultaneously increases the need to find measures that accurately assess teacher effectiveness.

Of immediate importance is determining and understanding valid measures of teacher effectiveness, which contribute to a complete picture or broader understanding of individual and collective effectiveness. The educational research and evaluation community must be able to credibly and fairly assess teacher effectiveness in both a formative way, to improve teaching (teacher quality, content knowledge, pedagogical practices, etc.), and a summative way, for decision making (e.g. selection of appropriate professional development, tenure decisions, student and teacher placement decisions, etc.).

This study explores the sensitivity of measures to detecting differences (within-group and between-group), relationships between teacher effectiveness constructs and measures, and the extent to which depictions of teachers vary across the different measures of effectiveness. Findings from this study help the educational research community to better understand estimates of effectiveness provided by different measures. Only when we truly understand the multiple aspects of teacher effectiveness can we fully ascertain teacher quality for formative and summative purposes. Ultimately, findings can be used to more meaningfully select measures to include in a comprehensive teacher evaluation system.

Conceptual Framework

By the 1980's, after the publication of the report, *A Nation at Risk*, which highlighted the inadequate education system in the United States, the imperative of raising student achievement and improving teacher quality was in full swing. In 1986, the Carnegie Task Force on Teaching as a Profession issued a report, *A Nation Prepared: Teachers for the 21st Century*, which paved the way for the establishment of The National Board for Professional Teaching Standards (NBPTS). In 1987, NBPTS was founded on the mission of improving teaching quality and learning (NBPTS, 2002). With the NBPTS's policy statement: *What teachers know and should be able to do* (1989 – First Edition), effectiveness standards were set for all teachers. NBPTS has gone on to publish advance standards for 24 different content areas across grade levels, which are widely regarded as the highest standards in the teaching profession. Although there is consensus that quality teachers are critical to student success and much work has been done to identify elements of effective teaching, schools have not done enough to evaluate teachers accurately, or used evaluative information to improve educational quality (The New Teacher Project, 2010).

The reauthorization of the *Elementary and Secondary Education Act*, passed by Congress in December 2001, better known as the *No Child Left Behind Act* (NCLB), of 2002 (P.L. 107-110), has created tremendous pressure on states receiving Title I dollars to be more accountable for the effects of educational services provided, namely student outcomes. NCLB recognizes the link between teacher quality and student achievement and mandates that all students are taught by, "highly qualified" professionals, for the first time in history, establishing national criteria for teachers (NCLB, Title II, Part A). The act requires that, at a minimum, teachers have a Bachelor's degree, are fully certified (or licensed) by the state in which they teach, and that they demonstrate subject area knowledge for every core subject that they teach. One of the ways NCLB aims to close the achievement gap is through providing all children with highly qualified teachers. The goal is to increase the number of highly qualified teachers and distribute them more equitably, especially in lower-performing schools, with higher percentages of minority and socioeconomically disadvantaged students. While a "qualified teacher" is one important aspect of teacher quality, there is more involved in ensuring equal access to effective teaching.

In the era of standards-based reform and accountability, districts are more and more concerned with improving educational outcomes (teachers' knowledge, skills, practices, and student learning and achievement) (Puma and Raphael, 2001). Intensive professional development opportunities are seen as one way to achieve intended educational outcomes (Puma and Raphael, 2001). Professional development (PD), in general, is one way teachers across the country are supported in increasing effectiveness. Public schools are spending $20 billion annually on professional development activities (NCES, 2008). California alone spent $1,382,000 in 2008 on professional development and teachers report spending between 40 to 120 hours a year engaged in various forms of PD (CPEC, 2009). Numerous studies show that effective professional learning can dramatically improve student learning and achievement legitimizing the huge investments in PD efforts (Tom Torlakson's Task Force on Educator Excellence, 2012).

Although funding for professional development has declined in recent years, it is still recognized that "providing a high-quality teacher in every classroom and effective education leaders in our public school systems is imperative" (Tom Torlakson's Task Force on Educator Excellence, 2012, p. 7.). "Expert teachers and leaders are perhaps the most important resource for improving student learning" pointing to the persistent need to invest substantially in teacher quality (Tom Torlakson's Task Force on Educator Excellence, 2012, p. 7). While billions of dollars are spent on professional development programs, valid and reliable measures of sustained professional development effects on teachers are lacking. Staffing, believed to be one of the most important elements in effective schooling, needs to be evaluated in valid and reliable ways. With the huge push towards improving teacher quality, we must have the means (valid and reliable measures of teacher effectiveness sensitive to detecting change) to assess progress.

A key tenet of *No Child Left Behind* is raising student achievement. NCLB has placed a tremendous amount of pressure on teachers to be held accountable for raising student achievement. One of the specific ways in which teachers are supported in this effort is through the Improving Teacher Quality (ITQ) State Grants program, provided by the U.S. Department of Education (under Teacher and Principal Training and Recruiting). The purpose of the program is to increase academic achievement by improving teacher and principal quality (U.S. Dept. of

Ed.). Title II, Part A funds are used to support various Professional Development (PD) programs across the country. The U.S. Department of Education, state governments, school districts, universities, and private educational programs are forming partnerships to deliver large scale and individualized PD where needed. This trend is especially apparent in lower performing school districts and schools in which raising the quality of teachers is seen as vitally important. Under the current legislation, there is no requirement that Title II funds be linked to assessment of teacher effectiveness, or that changes in effectiveness be recorded (The New Teacher Project, 2010). Although there is significant support in providing professional development, assessing the effects of these programs has not been prioritized and remains difficult.

One of the more recent educational initiatives funded by the federal government is the Race to the Top Fund (R2T). Race to the Top is a $4.3 billion educational reform program enacted as a part of the American Recovery and Reinvestment Act (ARRA), which awards grants to states through a competitive application process (The New Teacher Project, 2009). The U.S. Department of Education (USDE) awarded funds in early 2010 and Fall 2010 to states under four assurance areas: adopt common standards and assessments and make a plan for instituting them (standards and assessments); create the infrastructure to support statewide longitudinal data systems which link student and teacher data to support instruction (data systems to support instruction); differentiate teachers and principals based on effectiveness and incorporate the assessments into human capital policies and decisions (great teachers and leaders); and build the authority and framework to intervene with struggling schools and support high-quality charter schools (turning around struggling schools) (The New Teacher Project, 2009). R2T places great emphasis on improving teacher effectiveness by identifying differences in teachers in terms of impact on student achievement and connecting policies and decisions to assessments of effectiveness (The New Teacher Project, 2009).

R2T not only places a strong focus on assessing and maximizing teacher effectiveness, but also offers a prescriptive way to evaluate effectiveness. All funded states must implement a multi-measure and summative (4-point) rating category evaluation system (highly effective, effective, developing, and ineffective). The goal of the new evaluation system is to improve student achievement by: optimizing new teacher supply by hiring from programs proven effective; boost effectiveness of all teachers by targeted professional development; retain and leverage the most effective teachers; prioritize effective teachers for high-need students; and improve or exit persistently ineffective teachers (The New Teacher Project, 2009). States across the country are being inspired by R2T to revamp educational evaluation.

California, although not a winner of early Race to the Top funding, is still engaging in efforts to restructure the educational evaluation system. In response to the April 28, 2009 Board Motion (Quality Leadership and Teaching to Ensure a World Class Education for All), the Los Angeles Unified School District (LAUSD) formed the Teacher Effectiveness Task Force "to develop recommendations for enhancing the ways in which the district ensures that the most effective teachers, administrators and support personnel work with our students everyday" (LAUSD, 2010, p. 7). Composed of teachers, administrators, district leaders, parents, community representatives, and various other private and public sector stakeholders, the task force has published recommendations across (five sub-committee) areas of: evaluation; tenure; differentiated compensation; support mechanisms; and legislation. In 2010, The Teacher Evaluation Sub-Committee offered four recommendations for teacher evaluation in LAUSD:

> "Recommendation 1: Teacher evaluations should include multiple measures or data points." "Measures should include the following – teacher practice; student outcomes; parent and student feedback; collaborative/contribution to the school community; and self-evaluation." Currently there are diverging perspectives on weighting measures and data collection procedures. (LAUSD, 2010, p. 12)
>
> "Recommendation 2: Increase the number of rating categories (gradations) available." This is moving away from the STULL form categories of "meets standard performance" or "below standard performance." (LASD, 2010, p. 13)
>
> "Recommendation 3: Evaluations should have real ramifications." Both rewards and consequences are proposed. Currently there are diverging perspectives on how positive or negative consequences should be tied to evaluation results. (LAUSD, 2010, p. 13)
>
> "Recommendation 4: Professional development and support must be tied to feedback from evaluation" (LAUSD, 2010, p. 13).

On April, 27, 2010, LAUSD published *Immediate Action Steps Addressing the Recommendations of the Teacher Effectiveness Task Force*, declaring: "The single most important issue for this school district is to ensure every classroom is led by an effective teacher, every school is lead by an outstanding leader, and there is a team of excellent support personnel" (p. 1). By Fall 2016, LAUSD aims to have an effective principal at every school and an effective teacher in every classroom (LAUSD, 2010). One of the ways LAUSD is going to accomplish this is by developing a "multi-measure evaluation system that focuses on how [teacher] efforts support student learning" (LAUSD, 2010, p. 1). The district's *Three-year Strategic Plan*, indicates 2010-2011 as a prototyping year, where much is to be researched and discussed. 2011-2012 as a piloting year, where some of the proposed changes are tried in a small sample of schools, and 2012-2013 as the year that they take parts of the new evaluation system to scale (LAUSD, 2010).

In support of this effort, LAUSD has published new “Proposed LAUSD Teaching Standards” which are integral to effective teaching. The standards are designed to reflect the complexity of teaching and allow for deep understanding of professional standards. The framework is based on the research on effective teaching done by Charlotte Danielson (2007). Danielson’s model covers four major domains of effective teaching: planning and preparation; the classroom environment; instruction; and professional responsibilities. The goal is to clearly define standards, components, and elements so that teachers and evaluators understand the new expectations.

LAUSD indicated that one of the “key considerations for implementation,” was “the importance of choosing the right measures” (LAUSD, 2010, p. 23). The Teacher Effectiveness Task Force recommended “seeking guidance from the research community and vetting various approaches with key stakeholder groups” (LAUSD, 2010, p. 23). In response to the Teacher Effectiveness Task Force report, the Los Angeles Educational Research Consortium also encouraged the district to “Try out different measures of teacher effectiveness before settling on the “right” set of measures” (LAUSD, 2010, p. 36). With the plethora of teacher effectiveness measures available, and the deep complexities of teaching, it is important to build a system that measures what it values and values what it measures, in essence gets it right.

In Teacher Evaluation 2.0, The New Teacher Project states:

> “Evaluations should provide all teachers with regular feedback that helps them grow as professionals, no matter how long they have been in the classroom. Evaluation should give schools the information they need to build the strongest possible instructional teams, and help districts hold school leaders accountable for supporting each teacher’s development. Most importantly, they should focus everyone in a school system, from teachers to the superintendent, on what matters most: keeping every student on track to graduate from high school ready for success in college or a career.” (TNTP, 2010, p. 1)

In light of our current national and local (LAUSD) spotlight on revamping the educational evaluation system, it is a perfect time to study multiple ways in which to evaluate teacher effectiveness. Right now we are faced with building a rigorous, fair, and credible teacher evaluation system centered on student outcomes (The New Teacher Project, 2010). "The next several years represent a golden opportunity to create better systems that meet the needs of schools and the professionals that work in them" (The New Teacher Project, 2010, p. 1). Understanding what unique information different measures can offer about teacher effectiveness is absolutely critical. New evaluation systems are being built and informed decisions need to be made on which measures to include.

Study Purpose and Research Questions

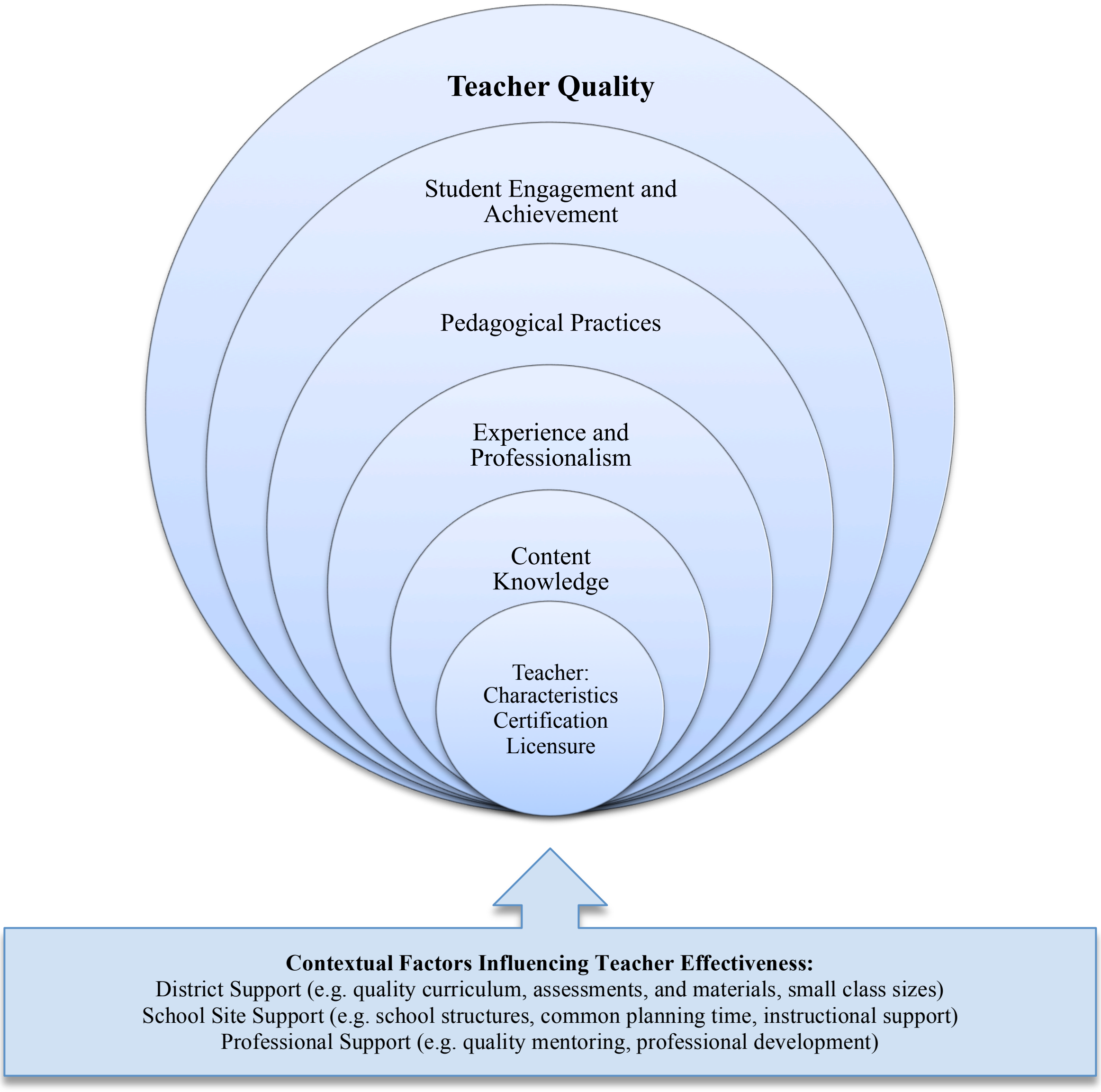
This study helps people involved in all stages and aspects of educational evaluation, especially those involved in the creation of new comprehensive teacher evaluation systems (CTESs). This study seeks to create an understanding of what three distinct measures of teacher effectiveness can contribute to an overall understanding of teacher quality. Articulations of teacher quality are directly impacted by the measures involved in ascertaining effectiveness making it imperative that measures are chosen and used appropriately. Through close examination of a teacher survey, expert assessment, and classroom observation protocol, each measure's suitability for inclusion in a CTES is determined.

This study takes place within the context of a state-funded Improving Teacher Quality Program (ITQP) designed to increase teacher effectiveness. This proves to be a perfect setting to investigate the measures used in assessing effectiveness because of the expected changes as a result of significant participation in high quality PD. The ITQP provider (UCLA Center X) is named by Tom Torlakson's Task Force on Educator Excellence, in the report *Greatness by Design* (2012), as a model PD partnership, as is the focus on National Board Certification activities, which ITQP incorporated into year three PD activities. This context provides fertile ground for assessing each measure's sensitivity to assessing effectiveness, and capturing changes in effectiveness, extremely important capabilities needed in the push to improve teacher quality in our nation.

Teacher quality is viewed as one large construct encompassing multiple elements of teacher effectiveness (e.g. certification, content knowledge, reflective practice, rapport with students, student achievement, etc.). Each element (construct/sub-construct) or area of investigation potentially contributes a unique depiction of teacher effectiveness. As I conceptualize it, in Figure 1, estimates of effectiveness, taken together, form a complete picture of teacher quality. Figure 1, contains examples of the many elements of teacher effectiveness contained within the construct of teacher quality, including, inputs, processes, and outputs. Estimates of teacher quality are influenced by the specific area(s) of investigation targeted to draw conclusions. For example, one could look at student achievement alone to garner an understanding of teacher quality, or, a combination of any of the elements housed within the larger construct to produce a more rounded view. The more credible evidence used, the more valid the estimate of teacher quality.

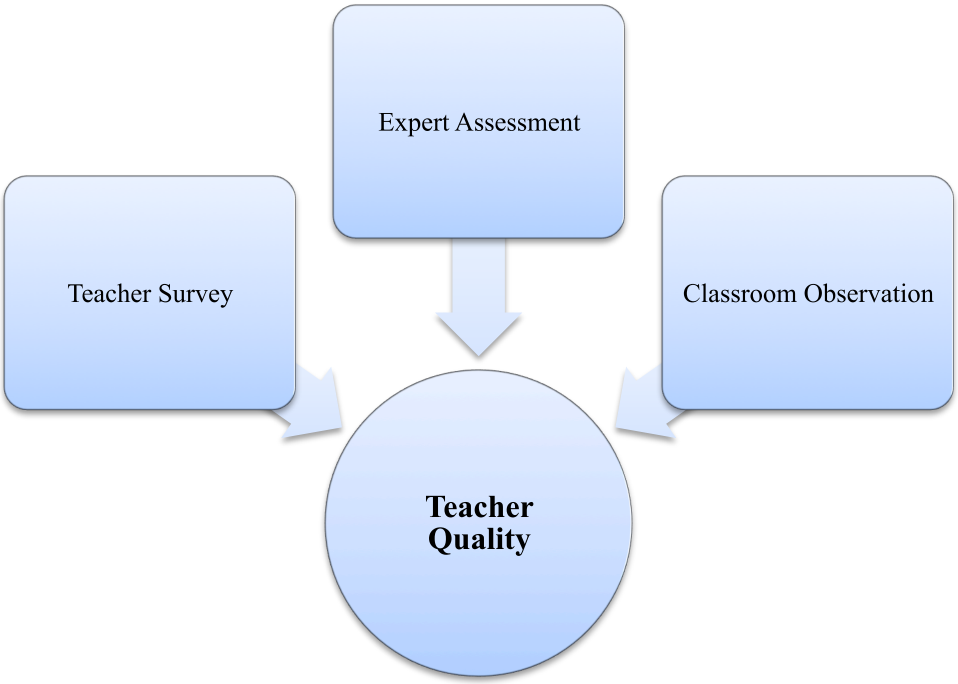
A factor that then influences the depiction of teacher quality is the measure or measures used and the sub-constructs contained within. Looking at individual measures of teacher effectiveness is important because it is the combination of the estimates of teacher effectiveness that give an estimate of teacher quality. As seen in recent educational legislation and debates, teacher quality has huge implications for equity and access for students, student achievement,

Figure 1.
Illustration of Teacher Quality with Example Indicators of Teacher Effectiveness, including Influential Contextual Factors


Teacher Quality
Student Engagement and Achievement
Pedagogical Practices
Experience and Professionalism
Content Knowledge
Teacher: Characteristics Certification Licensure
Contextual Factors Influencing Teacher Effectiveness:
District Support (e.g. quality curriculum, assessments, and materials, small class sizes)
School Site Support (e.g. school structures, common planning time, instructional support)
Professional Support (e.g. quality mentoring, professional development)

teacher pay and tenure, etc. When inferences about teacher quality depend upon the assessment(s) of teacher effectiveness, it becomes extremely important to understand what each measure contributes to defining teacher effectiveness and how estimates of effectiveness vary across measures. The overall purpose of this study is to explore the sensitivity and predictive power of various estimates of teacher effectiveness and the extent to which depictions of teachers hold across the different measures. Figure 2 illustrates the measures of teacher effectiveness used to make inferences about teacher quality in this study[1].

Figure 2.
Map of Teacher Quality in this Study, Defined by a Combination of Teacher Effectiveness Measures



Another factor that I believe influences the depiction of teacher quality is the unit of analysis contained within the measure(s) of teacher effectiveness. One common element used to assess teacher effectiveness is student achievement on standardized tests. When students are used as the unit of analysis, teachers are no longer directly assessed. Students are one degree

[1] The original vision for this study included CST scores for 8th grade teachers as a measure of teacher effectiveness by way of student achievement. Unfortunately, due to a legal stalemate between the school district and the institute of higher learning, student level CST data was never made available to the Improving Teacher Quality Program (ITQP) study. Because all data for this study is pulled from ITQP, student level CST scores are not available for analysis.

removed from teachers and therefore give an indirect measure of teacher effectiveness. Although it is important to consider the outcome of teaching (i.e. student achievement) basing teacher evaluation on student performance introduces a level of "noise" or bias into findings (no matter the attempts to control for such biases). When teacher quality is largely based on student achievement, an in-depth view of teacher quality is lacking. Considering the high-stakes placed on teacher evaluation, anything less than an in-depth, or complete view, of teacher quality should be rejected.

It is important to understand how different measures of teacher effectiveness influence understanding of teacher quality. We need to know which measure(s) add(s) something unique to the understanding of teacher effectiveness, and what combination of measures provides a well-rounded view of teacher quality. This study takes a step in understanding which measures should be contained in a comprehensive teacher evaluation system, aimed a providing a complete assessment of teacher quality.

There is surprisingly little research on how individual teachers perform across different measures of teacher effectiveness. Many studies explore teacher effectiveness based on one or two measures (Bakke, 1999; Brown, 2004; Carter, 2008; Forte, 1999; Hill, 2002). It is difficult to make inferences about different measures across studies because the populations of teachers differ. Much insight can be gained in collecting in-depth longitudinal data on teacher effectiveness for a select group of teachers. Having data on multiple dimensions of individual teacher effectiveness for select group of teachers allows for a more complex discussion of the actual measures used to generate data for the analyses. This study helps generate an understanding of how measures of teacher effectiveness compare to one another by looking at the depictions of teachers across teacher effectiveness constructs and measures.

There are also studies that address the extent to which teacher effectiveness grows as a result of exposure to an intervention (Forte, 1999; Gargani, 2009; Lansman, 2006). These studies typically use one or two measures to assess an intervention's impact on raising teacher effectiveness. One difficulty is that each study uses different measures with different populations of teachers, making it impossible to make connections across studies and measures. Having in-depth longitudinal effectiveness data available on a select group of teachers provides an understanding of which measure(s) is/are sensitive to capturing differences (both between-and-within-groups) in teacher effectiveness. This is important because with such emphasis placed on assessing and improving teacher quality, measures of teacher effectiveness must be sensitive enough to detect changes. Assuming that the goal to improve teacher quality persists, there must be a way to capture progress. This study aids in understanding the sensitivity of measures to detecting differences within-groups over time and between-groups, and considers the predictive power of teacher characteristics and a professional development intervention in determining scores on teacher effectiveness constructs.

This study uses pre-existing data supplied from a recent study of an Improving Teacher Quality program (ITQP). ITQP was a professional development program located in two Partnership for Los Angeles Schools (PLAS), located within Los Angeles Unified School District (LAUSD). Using data that was collected from surveying, assessing, and observing urban middle school science and social studies teachers, this study addresses the following sets of research questions:

1. What teacher effectiveness measure(s) exhibit(s) the greatest level(s) of sensitivity to detecting differences?

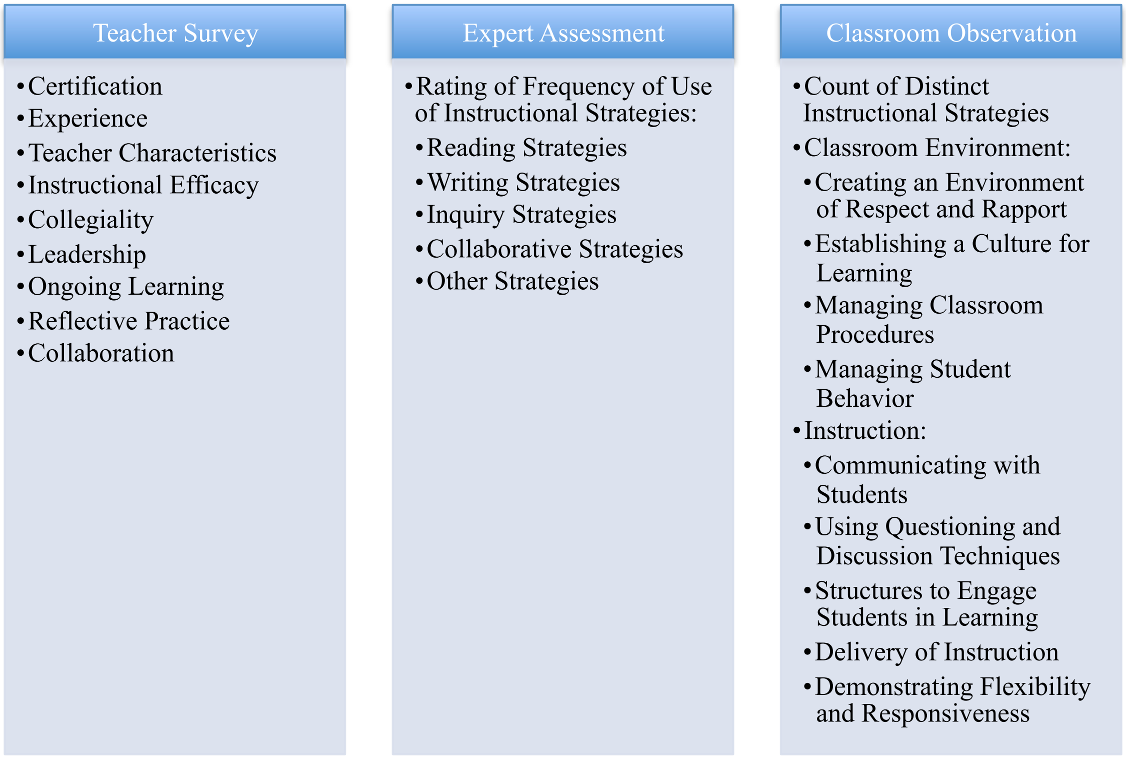
   a. How does each measure capture changes in teacher effectiveness over time?

    b. How does each measure capture group differences between low, moderate, and high (ITQP) participation teachers?

2. What are the relationships between measures of teacher effectiveness and teacher characteristics?
    a. What is the strength and direction of association between variables?
    b. In what ways are teacher characteristics predictive of teacher effectiveness?
    c. In what ways are academic PD participation and/or total strategy use predictive of teacher effectiveness?
3. In what ways are individual teachers depicted similarly and differently across different measures of teacher effectiveness?
    a. In what ways do these patterns hold across teacher effectiveness constructs?
    b. In what ways do these patterns hold across measures?

Figure 3, below, lists the effectiveness elements and sub-constructs contained within each of the measures of teacher effectiveness. Each of these listed elements and sub-constructs is used to answer the research questions stated above.

Figure 3.
Map of Elements and Constructs Contained within Teacher Effectiveness Measures Used in this Study



Teacher Survey

- Certification
- Experience
- Teacher Characteristics
- Instructional Efficacy
- Collegiality
- Leadership
- Ongoing Learning
- Reflective Practice
- Collaboration

Expert Assessment

- Rating of Frequency of Use of Instructional Strategies:
  - Reading Strategies
  - Writing Strategies
  - Inquiry Strategies
  - Collaborative Strategies
  - Other Strategies

Classroom Observation

- Count of Distinct Instructional Strategies
- Classroom Environment:
  - Creating an Environment of Respect and Rapport
  - Establishing a Culture for Learning
  - Managing Classroom Procedures
  - Managing Student Behavior
- Instruction:
  - Communicating with Students
  - Using Questioning and Discussion Techniques
  - Structures to Engage Students in Learning
  - Delivery of Instruction
  - Demonstrating Flexibility and Responsiveness

Study Significance and Implications

This study makes contributions to three primary areas. This study:

1. Increases the knowledge around and understanding of measures of teacher effectiveness. This is important because evaluations of teacher effectiveness and subsequent evaluations of overall teacher quality can be used to make vital decisions about a teacher's career (hiring, firing, tenure, pay, professional development, promotion, etc.). If teacher quality is directly dependent upon the depictions of effectiveness, it is vitally important that we thoroughly understand the measures being used to assess effectiveness. How estimates of effectiveness vary across measures helps us to better understand individual estimates of teacher effectiveness and larger inferences about teacher quality. Because of the high

stakes involved in teacher evaluation, assessment of teacher effectiveness must be valid, reliable, and complete.

2. Helps those involved in educational research and educational program evaluation better select measures of teacher effectiveness to include in research or evaluation studies. With the large push towards portfolio evaluation of teachers it is important that the "right" measures are selected for inclusion in models determining teacher quality and that we know what can be inferred from selecting those measures. Because so much is at stake, for both teachers and students, accurate teacher effectiveness measures and in-depth understanding of those measures are needed. Researchers and evaluators must fully understand not only the dimensions of teacher effectiveness, but also how to measure teacher effectiveness in order to create an accurate and holistic view of teacher quality.

3. Helps those involved in improving teacher quality programs and interventions (or similar programs) to better capture changes in teacher effectiveness and overall teacher quality. No Child Left Behind has deeply impacted the educational system, directing many resources towards improving teacher quality programs. These programs typically offer teachers a professional development intervention spanning anywhere from a couple of days to several years. Selecting measures that are sensitive enough to detect changes in teacher effectiveness increases the understanding of improving teacher quality programmatic effects, which can be used for program and policy decision-making.

Manuscript Organization

In the following four chapters of the manuscript, I provide context for the study and summarize study procedures and findings. Chapter 2 is comprised of a review of the relevant literature in effective teaching (both in general and more specifically for middle school science and social studies/history), effective professional development, and measuring teacher effectiveness. In Chapter 3, I describe the study's methods, including an overview of the study participants, instrumentation, and analysis. Additionally, I describe the context of the Improving Teacher Quality Program (ITQP), including, the site and participants, and corresponding evaluation. Chapter 4 is the summary of research findings in terms of the research questions stated above. Lastly, in Chapter 5, I draw conclusions from the study results and address implications for the field of teacher evaluation. The final chapter also includes a discussion of study limitations and suggestions for future research.

CHAPTER 2: Review of Relevant Literature

This in-depth study of measures of teacher effectiveness involves mainstream science and social studies-history teachers at two middle schools in a large, urban school district (LAUSD). This study seeks to create an understanding of 3 distinct measures of teacher effectiveness that involve surveying, assessing, and observing teaching practice. In order to assess teacher effectiveness one must first start with an operational definition of effective teaching. Moreover, to assess science and social studies-history teachers' effectiveness, one must understand effective teaching specific to that content area. In order to examine the specific measures used to evaluate effectiveness, one must have a firm grasp on measuring effective teaching. Given that this study takes place within the context of a professional development program, geared at improving teacher effectiveness, this chapter also addresses components of effective professional development. This chapter reviews some of the key literature in the areas of effective teaching, standards for early science and social studies education, effective professional development, and measuring teacher effectiveness.

Effective Teaching

There is strong consensus that teaching quality is key to student success (Darling-Hammond, 2000 and 2011; Ferguson, 1991; Kemp & Hall, 1992; Rice, 2003; Rivers & Sanders, 2002; Sanders & Rivers, 1996). Consistently, studies have shown that "good teaching" and "teachers matter" (Haycock, 1998, p. 1). The quality of one's teacher is directly related to the quality of one's learning and achievement. A study by Hanushek, et al. (1992), estimates that the individual teacher effect is as high as 7% of the total student achievement effect. Another study by Gordon, et al. (2006), finds that the difference in effect between a top quartile teacher and a bottom quartile teacher is as high as 10 percentile points. In 1991, Ronald Ferguson found that as high as 40% of the variation in student achievement is attributed to the quality of the instructor. Even more staggering, Dr. Darling-Hammond found that teacher qualification accounted for 90% of the variation in student achievement in reading and math (2000).

The differences in effect on student achievement between a highly effective teacher and an ineffective teacher are both statistically and practically significant. Given the achievement gap nationally between African-American students and White students is around 34 percentile points (Gordon, et al., 2008), the cumulative effect of highly effective versus ineffective teachers has huge implications. If all students could have access to highly effective teachers for several consecutive years, the achievement gap may largely disappear (Gordon et al., 2008; Haycock, 1998).

What constitutes effective teaching is something that has been widely discussed in the field of education for decades. Prodigious amounts of research have been conducted on identifying the teacher characteristics and best practices that result in improved student learning and increases in student achievement. Although there is a plethora of studies on effective teaching, there is yet to be widely held consensus on what effective teaching entails. Part of this is due to the fact that teacher effectiveness is such a broad concept and most studies narrow in on a small aspect of teacher effectiveness. Teacher effectiveness essentially covers three pertinent areas: inputs (characteristics, certification, content knowledge, etc.); processes (classroom ecology, curricular resources and materials, pedagogical practices, teacher-student interactions, etc.); and outputs (student engagement, parental involvement, student learning, student achievement, etc.) (Goe, Bell, & Little, 2008).

State licensing systems focus on the inputs of effectiveness, setting minimum standards for certification for beginning teachers. These are the same standards referenced by NCLB when referring to “highly qualified teachers.” Although these standards establish entry-level requirements, they do not adequately cover advanced standards for experienced teachers. The process or output components of teacher effectiveness are completely neglected by NCLBs standards for qualified teachers.

A more recent definition of teacher effectiveness, supplied in the Federal Race to the Top (R2T) initiative focuses solely on a teacher’s effect on students, stating that an effective teacher is one whose “students achieve acceptable rates (e.g. at least one grade level in an academic year) of student growth” (USDOE R2T website, p. 7). R2T further defines a teacher as highly effective when “students achieve high rates (e.g. one and one-half grade levels in an academic year) of student growth” (USDOE R2T website, p. 7). To be eligible for R2T funding, states, LEAs, and schools must establish multiple measure teacher evaluation systems that evaluate teacher effectiveness, in significant part, by student growth. Supplemental measures are also included and range in modalities (multiple observation-based assessments of teacher performance, surveys, student work, etc.). R2Ts definition of teacher effectiveness is not an adequate definition because it narrowly focuses on outputs, namely student achievement, ignoring inputs and processes (and other outputs of teaching).

An effective teacher is not equal to a “highly qualified” teacher or simply a teacher that has students with high gains in student achievement. Teaching is a complex endeavor that demands a definition of effectiveness that values and reflects that complexity. In 1989, The National Board for Professional Teaching Standards issued a policy statement, *What teachers* *know and should be able to do*, delineating the highest standards for effective teaching. The policy brief presents the values of the National Board.

> "The fundamental requirements for proficient teaching are relatively clear: a broad grounding in the liberal arts and sciences; knowledge of the subjects to be taught, or the skills to be developed, and of the curricular arrangements and materials that organize and embody that content; knowledge of general and subject-specific methods for teaching and for evaluating student learning; knowledge of students and human development; skills for effectively teaching students from racially, ethnically, and socioeconomically diverse backgrounds; and the skills, capacities and dispositions to employ such knowledge wisely in the interest of students…. Teaching ultimately requires judgment, improvisation, and conversation about means and ends. Human qualities, expert knowledge and skill, and professional commitment together compose excellence in this craft" (p. 2).

The policy brief also established *Five Core Propositions* that describe the knowledge, skills, abilities, and commitments of a teacher who effectively enhances student learning. These five tenets guide professionalism in schools and serve as the basis (in combination with individual standards for each content area) for all National Board Certification. On pages 3-4 and 8-20, the *Five Core Propositions* are described as follows:

1. "*Teachers are committed to students and their learning* – Teachers recognize individual differences in their students and adjust their practice accordingly. Teachers have an understanding of how students develop and learn. Teachers treat students equitably. Teachers' mission extends beyond developing the cognitive capacity of their students."

2. "*Teachers know the subjects that they teach and how to teach those subjects to students* – Teachers appreciate how knowledge in their subjects is created, organized and linked to other disciplines. Teachers command specialized knowledge of how to convey a subject to students. Teachers generate multiple paths to knowledge."

3. "*Teachers are responsible for managing and monitoring student learning* – Teachers call on multiple methods to meet their goals. Teachers orchestrate learning in a group setting. Teachers place a premium on student engagement. Teachers regularly assess student progress. Teachers are mindful of their principal objectives."

4. *Teachers think systematically about their practice and learn from experience* – Teachers are continually making difficult choices that test their judgment. Teachers

seek the advice of others and draw on Educational research and scholarship to improve their practice."

5. *Teachers are members of learning communities* – Teachers contribute to school effectiveness by collaborating with other professionals. Teachers work collaboratively with parents. Teachers take advantage of community resources."

Laura Goe, Courtney Bell, and Olivia Little have done extensive research in the areas of effective teaching and measuring teacher effectiveness. In their article from 2008 (p. 8), they offer a five-point definition of teacher effectiveness, covering multiple components.

1. "Effective teachers have high expectations for all students and help students learn, as measured by value-added or other test-based growth measures, or by alternative measures."

2. "Effective teachers contribute to positive academic, attitudinal, and social outcomes for students such as regular attendance, on-time promotion to the next grade, on-time graduation, self-efficacy, and cooperative behavior."

3. "Effective teachers use diverse resources to plan and structure engaging learning opportunities; monitor student progress formatively, adapting instruction as needed; and evaluate learning using multiple sources of evidence."

4. "Effective teachers contribute to the development of classrooms and schools that value diversity and civic-mindedness."

5. Effective teachers collaborate with other teachers, administrators, parents, and education professionals to ensure student success, particularly the success of students with special needs and those at high risk of failure."

In February of 2011, Linda Darling-Hammond gave a talk at the (then) Annual California Post Secondary Education Commission's Research Director's Meeting (for which I was present). Framed within the context of ensuring effective teaching for every child, she described: effective teaching; ways in which we could better develop our teaching force; and the most promising ways to evaluate teacher effectiveness. According to Darling-Hammond (2011),

"Effective Teachers…

- engage students in active learning.
- create intellectually ambitious tasks.
- use a variety of teaching strategies.
- access student learning continuously and adapt teaching to student needs.
- create effective scaffolds and supports.
- provide clear standards, constant feedback, and opportunities for revising work.
- develop and effectively manage a collaborative classroom in which all students have membership."

Dr. Darling-Hammond (2011) also discussed research that identifies the teacher components related to increases in student achievement. Citing Clotfelter, Ladd, & Vigdor (2008), she advocates for teachers with ample preparation, including "strong academic background, certification in the subject area taught, National Board Certification," and teaching "experience that is greater than 3 years." She also talked about how school site support and other factors can influence teacher effectiveness, including, the extent to which teachers receive quality mentoring, professional development, and other instructional support. She discussed how schools can be structured in ways that facilitate collaboration amongst teachers by giving common planning time and personalizing the structure of the school, and how district support, such as quality curriculum and assessments, other materials, and small class sizes influence teacher effectiveness (Darling-Hammond, 2011).

Dr. Darling-Hammond also touched on the idea that there are many other factors, beyond teacher quality, that contribute to student learning and achievement. The environment and resources available for schooling directly affect learning and achievement (Darling-Hammond, 2011). Factors such as: prior learning opportunities; home context; attendance; materials; curriculum quality; coherence and continuity; class size; etc. all contribute to student learning and achievement (Darling-Hammond, 2011). Teacher effectiveness can be mitigated or enhanced based on many of the aforementioned factors.

Other studies further expand on elements of effective teaching referring to various intrapersonal (mission, investment, focus); interpersonal (empathy, rapport, listening); and extrapersonal (individualized perception, drive, activation, innovation, gestalt, leadership) traits and qualities (Bakke, 2004; Gallup Inc., 1978). Some studies look at the presence of even additional aspects of teacher effectiveness: enthusiasm; organization; group interaction; individual rapport; breadth of coverage; examinations/grading; assignments; and workload/difficulty (Brown, 2004). Findings from other research suggests effective teachers: clearly articulate classroom norms and run orderly classrooms (Kemp & Hall, 1992); keep students on task and highly engaged in material (Kemp & Hall, 1992; Taylor, Pearson, & Walpole, 1999); begin lessons with review (Kemp & Hall, 1992); employ systematic teaching strategies (Kemp & Hall, 1992); spend significant amounts of time working in small groups (Taylor, Pearson, & Walpole, 1999); differentiate instruction (Kemp & Hall, 1992); offer multiple opportunities to apply learning (Kemp & Hall, 1992); regularly check for understanding (Kemp & Hall, 1992); offer systematic feedback on student work (Kemp & Hall, 1992); and communicate regularly with parents (Taylor, Pearson, & Walpole, 1999).

When looking at the research on teacher effectiveness, it is clear that teaching is a complex profession, with many factors and components that interact and influence teacher effectiveness. Most descriptions of effective teaching cover aspects such as teacher qualifications, "organization of the subject matter, skills in instruction, and personal qualities and attitudes that are useful when working with students" (Edwards, 2005, p. 50). While there are multiple interpretations of effective teaching, it remains undisputed that teacher quality is a critical determinant of student achievement. This study takes a broad view of teacher effectiveness, focusing on the inputs, processes, and outputs of teaching.

Standards for Early Science Education

While the definitions of effective teaching outlined above are extensive in nature, they apply to the teaching profession as a whole. In addition to the more general definitions of teacher effectiveness, each content area (and grade level) has additional sets of expertise that apply to effective teaching. In order to assess science teachers' effectiveness in this study, it is important to understand standards and expectations that are specific to that field. In this section, I elaborate on the definition of effective teaching specifically for teachers at the middle school level in the science content area.

Science is an important core subject area through the elementary, middle and high school levels (and beyond). Science pervades our everyday lives from our living planet to the technology we use. According to the National Board for Professional Teaching Standards, through a mix of craft, artistry, proficiency, and understanding, science teachers help students understand and make sense of the world that we live in (2003). More specifically, …

> "Accomplished science educators' primary goal is to develop scientifically literate students by teaching them to think like scientists, both in science class and their everyday lives. Science educators set out to instill in students a never-ending curiosity about the world and to develop in them the skills necessary to investigate their questions. They challenge students to explore unanswerable questions, test hypotheses, construct and revise models, and question innovations. They make science exciting while ensuring that students are learning." (NBPTS, 2003, p. 1)

The National Board for Professional Teaching Standards (NBPTS) publishes field-specific standards for accomplished teaching, grounded in the *Five Core Prepositions*, articulated in section II-A. Each set of field-specific standards represents a consensus of accomplished professionals in the field on the attributes of practice that distinguish accomplished teaching in that field (NBPTS, 2010). The standards are defined according to the developmental level of the students and by content area. The standards set forth by NBPTS are regarded as the highest in the field and are used as the basis of field-specific effective teaching for this study. Effective science teachers integrate content guidelines, established by the state in which they are teaching, with the standards put forth by NBPTS. Each of the standards is expressed in terms of observable teacher practice and impact on students. There are thirteen science standards for ages 11-15, covering four broad areas of effective practice. Each standard is detailed with performance criteria and descriptive examples of observable teaching behaviors (not included here). The science standards detailed in *NBPTS Early Adolescence Science Standards: For Teachers of Students Ages 11-15* (2003), are described as follows:

Preparing the Way for Productive Student Learning

> Standard I: Understanding Early Adolescents – "Accomplished science teachers know the unique characteristics of their students and use this knowledge to determine students' understanding of science and to design and implement appropriate instruction to enhance student learning" (p.7).
>
> Standard II: Knowledge of Science – "Accomplished science teachers have a broad and current knowledge of science, along with in-depth knowledge of one of the subfields of science, on which they draw to set appropriate learning goals for their students" (p. 11).
>
> Standard III: Instructional Resources – "Accomplished science teachers are innovative in their ability to select, adapt, and create instructional resources, including print, technology, and community resources, to support active student explorations of science" (p. 17).

Establishing a Favorable Context for Student Learning

> Standard IV: Diversity, Equity, and Fairness – "Accomplished science teachers take steps to understand and value the diversity of all students, promote equity in the classroom and beyond, and uphold fairness in their daily interactions with all students" (p. 21).
>
> Standard V: Engagement – "Accomplished science teacher engage students in science through creative and innovative experiences" (p. 25).

Standard VI: Learning Environment – "Accomplished science teacher create stimulating and safe learning environments that foster high expectations for the success of all students and in which students experience the values inherent in the practice of science" (p. 29).

Advancing Student Learning

Standard VII: Understanding Science Pedagogy – "Accomplished science teachers understand and use a variety of instructional strategies to enhance student learning and help students make real-world connections from their scientific explorations" (p. 33).

Standard VIII: Science Inquiry – "Accomplished science teachers involve students in the process of inquiry that challenge students' thinking as they construct an understanding of nature and technology" (p. 39).

Standard IX: Contexts of Science – "Accomplished science teachers create opportunities for students to explore science in a variety of contexts, including its history, its reciprocal relationship with technology, and its impact on society" (p. 45).

Standard X: Assessment – "Accomplished science teachers employ a variety of assessment methods to obtain useful information about student learning and development, to guide instructional decisions, to report student progress, and to assist students in reflecting on their own learning" (p. 49).

Supporting Teaching and Student Learning

Standard XI: Family and Community Outreach – "Accomplished science teachers proactively work with families and communities to serve the interests of the students" (p. 55).

Standard XII: Professional Collaboration and Leadership – "Accomplished science teachers collaborate with colleagues and take leadership roles in their own educational community, as well as the larger community, to advance student learning" (p. 59).

Standard XIII: Reflective Practice – "Accomplished science teachers continually analyze, evaluate, and strengthen their practice in order to improve the quality of their students' learning experiences" (p. 63).

Standards for Early Social Studies-History Education

Just as it was important to define effective teaching for science teachers in this study, it is equally important to elaborate on the definition of effective teaching specifically for teachers in the social studies-history content area. In order to assess social studies-history teachers' effectiveness, it is important to understand the standards and expectations that are specific to that field. This section expands the overall definition of effective teaching for middle school social studies-history teachers.

Social studies-history is an important core subject area through the elementary, middle and high school levels (and beyond). The teaching of Social Studies-History is a particularly fascinating field. Social studies teachers have the opportunity to address difficult and controversial issues, encouraging students to engage in inquiry, critical reflection, evidence-based reasoning, and public deliberation (NBPTS, 2010).

> "Social studies-history teachers engage student in the most fascinating, exhilarating, maddening, and even confusing topics known to humanity: the origin and spread of scientific ideas; religions; ideologies; the nature of people, places, and environments; the meeting of cultures and the exchange of ideas; changes in love, marriage, and the family; the rise of democracies and dictatorships… Social studies-history teachers prepare students for participation in the public life of a democratic society… Each day's headlines are the content of the subject; every war or revolution, every social movement or election, every change in the economy or environment cries out for the contextualization that social studies-history can provide" (p. 13-14).

The National Board for Professional Teaching Standards (NBPTS) has also published field-specific standards for accomplished teaching for social studies-history. Again, the standards set forth by NBPTS are regarded as the highest in the field and are used as the basis of field-specific effective teaching for this study. Effective social studies-history teachers integrate content guidelines, established by the state in which they are teaching, with the standards put forth by NBPTS. Each of the standards is expressed in terms of observable teacher practice and impact on students. There are eight social studies-history standards, for ages 7-18, covering three broad areas of effective practice. Each standard is detailed with performance criteria and descriptive examples of observable teaching behaviors (not included here). The standards set forth in *Social Studies-History Standards: For Teachers of Students Ages 7-18+* (NBPTS, 2010), are described as follows:

Knowing Students, Purpose, and Content

> Standard I: Knowing Students – "Accomplished social studies-history teachers are knowledgeable about students as individuals and as members of families and communities and use their knowledge to strengthen relationships that increase student achievement. Teachers are also knowledgeable about students' development and their conceptualization of social studies-history" (p. 19).
>
> Standard II: Developing Social Understanding, Engagement, and Civic Identity – "Accomplished social studies-history teachers develop students' knowledge, skills, and attitudes necessary for social understanding and civic engagement and facilitate students' development as decision makers involved in public discourse and action at the local, national, or global levels" (p. 19).
>
> Standard III: Content – "Accomplished social studies-history teachers ground their teaching practice in a sound foundation of content knowledge" (p. 19).

Teaching in Context

> Standard IV: Instruction – "Accomplished social studies-history teachers recognize that excellent instruction depends on skilled organization and creative interweaving of curricula, varied instructional strategies, meaningful assessment, and supportive resources that engage students with content, provide meaningful and instructive feedback, and promote a love of learning" (p. 20).
>
> Standard V: Diversity – "Accomplished social studies-history teachers consider diversity a fundamental and deliberate component of teaching. Teachers recognize the importance of student diversity, equity in instruction, and pluralism in curriculum" (p. 20).
>
> Standard VI: Learning Environments – "Accomplished social studies-history teachers actively create and cultivate safe and dynamic learning environments characterized by respectful peer interactions, facilitation of multiple perspectives,

> and collaborative partnerships with families and with students' greater communities" (p. 20).

Developing as a Professional

> Standard VII: Professional Growth – "Accomplished social studies-history teachers pursue professional growth activities and experiences to develop themselves, their colleagues, schools, and districts, and to benefit the larger field of social studies-history education" (p. 20).
>
> Standard VIII: Reflection – "Accomplished social studies-history teachers engage in purposeful reflection as a systematic self examination of all aspects of teaching to extend knowledge, improve teaching, and refine their practice and their philosophy of education" (p. 20).

Effective Professional Development (PD)

This study takes place within the context of a state funded improving teacher quality professional development program. This context proves to be a perfect environment to test the sensitivity of different measures to capturing changes in teacher effectiveness due to the expectation for change in effectiveness as a result of significant participation in quality professional development. In order to provide further context for this study, it is important to create an understanding of what quality professional development entails. Just as there are multiple definitions of effective teaching, there are also multiple definitions of effective professional development. This section distinguishes between professional development and professional learning, addresses the various forms of PD delivery, and summarizes characteristics of effective professional development.

Professional development is designed to improve teacher effectiveness (by way of improving content knowledge, skills, and pedagogical practices) and ultimately student learning and achievement (CPEC, 2009; Wei, Darling-Hammond, Andree, Richardson, and Orphanos, 2009; Puma and Raphael, 2001; Florida Department of Education, 2010). Professional development is a label attached to activities led by experts designed to hone the knowledge and skills of teachers (Elmore, 2002; Wei et al., 2009). PD activities are considered formal activities that can be incorporated into the regular course of a teacher's work. As opposed to pre-service work, PD occurs once teachers are already on the job (Elmore, 2002; Wei et al., 2009). The value of PD is judged based on "what it contributes to the individual's capacity to improve the quality of instruction in the school and school system" (Elmore, 2002, p. 14). Interestingly, professional development on its own may or may not lead to professional learning (Tom Torlakson's Task Force on Educator Excellence, 2012).

Professional learning, may include professional development activities, but refers more generally to the ongoing, continuous development of educators (Florida Department of Education, 2010). Professional learning improves teaching practices and leads to increases in student achievement through both informal activities (e.g. collaborative study groups, learning communities, reflection on student work, action research, peer mentoring) and formal PD activities (Tom Torlakson's Task Force on Educator Excellence, 2012; Wei et al., 2009). According to Wei et al. (2009), professional learning can be a result of…

> "both formal professional development and other opportunities for professional learning – such as common planning time, shared opportunities to examine student work, or tools for self-reflection – that may occur outside the bounds of formal professional development events… Professional learning [is] a product of both externally-provided and job-embedded activities that increase teachers' knowledge and change their instructional practices in ways that support student learning. Thus, formal professional development represents a subset of the range of experiences that may result in professional learning" (p. 50).

Professional development comes in many forms. PD can be a formal structured activity, a job-embedded activity, or a new teacher induction program (Wei et al., 2009). PD activities are generally led by experts in the field and include formal consultations, workshops, summer institutes, seminars, courses, trainings, coaching, and mentoring (Elmore, 2002; Guskey and Yoon, 2009; Puma and Raphael, 2001). Activities can take place on-or-off school-sites and can last anywhere from one day to weeks or even years. Some times teachers get academic credit for PD, other times they may obtain certification in a specific area, fulfill district PD requirements, or become eligible for a stipend. Effective professional development is not restricted to any one form of PD, as characteristics of effective PD can be woven into many designs. The PD purposes, goals, and audience often help determine the appropriate form of PD.

While there are many characteristics of effective professional development reported in the literature, there is some general consensus on what effective PD entails. A study, conducted by Thomas Guskey (2003), analyzed 13 (well known and influential) lists of effective professional development characteristics. He found 21 distinct characteristics of effective professional development, with 6 characteristics appearing in over half of the lists: effective PD… (1) enhances teachers' content and pedagogic knowledge (92%); (2) provides sufficient time and other resources (77%); (3) promotes collegiality and collaboration (69%); (4) aligns with other reform initiatives (69%); (5) includes procedures for evaluation (62%); and (6) models high-quality instruction (54%) (Guskey, 2003).

Other studies echo similar characteristics, supporting the idea that effective PD is focused on the development of important content knowledge (Cohen and Hill, 1998; CPEC, 2009; Guskey and Yoon, 2009; Florida Department of Education, 2010; Shulman, 1986; Tom Torlakson's Task Force on Educator Excellence, 2012; Wei et al., 2009). Additionally, studies stress the importance of PD being sustained over time so that teachers can truly deepen their understanding of concepts and integrate new ideas and practices into their classrooms (CPEC, 2009; Elmore, 2002; Guskey and Yoon, 2009; Florida Department of Education, 2010; Killion, 2006; Puma and Raphael, 2001; Tom Torlakson's Task Force on Educator Excellence, 2012). In Guskey and Yoon's 2009 study, they found that teachers needed a minimum of 30 hours of PD contact hours to show positive effects on instruction. A study by Corcoran, Shields, and Zucker (1998) found that a minimum of 100 hours of contact time was needed for a PD program to have its intended effect on instruction. What is clear across studies is that a substantial allotment of time devoted to PD is integral to PD effectiveness. Research also supports that effective PD is collaborative in nature, focusing on improving schools and school systems rather than individual teachers (including development of curriculum and assessments, and lesson planning) (CPEC, 2009; Elmore, 2002; Tom Torlakson's Task Force on Educator Excellence, 2012). One important thing to note is that although the above mentioned characteristics are important, they must be (e.g. time and collaboration) implemented in a well organized, structured, and purposeful manner to achieve intended PD goals (Guskey, 2003). Since PD is specifically designed to improve teaching and student learning, research also suggests that it should be evaluated continuously on the extent to which it is achieving intended goals (Elmore, 2002).

Additionally, Guskey (2003) found 8 characteristics of effective professional development that appeared in at least a third (roughly) of the well-known and influential lists: effective PD… (7) is school or site based (46%); (8) is based on teachers' identified needs (46%); (9) focuses on individual and organizational improvement (46%); (10) is driven by analysis of student learning data (46%); (11) includes follow-up and support (38%); (12) is ongoing and embedded (38%); (13) is based on best available research evidence (31%); and (14) builds leadership capacity (31%).

The consensus view of effective PD clearly points to a follow-up component that helps support teachers in the implementation of new ideas and practices into their classrooms (CPEC, 2009; Elmore, 2002; Guskey and Yoon, 2009; Killion, 2006; Wei et al., 2009). Effective PD also tends to be embedded in teachers' own work and connected to the content and pedagogy questions that teachers are asking, as well as the difficulties students are encountering in real classrooms (Elmore, 2002; Tom Torlakson's Task Force on Educator Excellence, 2012). When professional development can take place in job-embedded activities, the potential for impact is greatest (Tom Torlakson's Task Force on Educator Excellence, 2012). Effective PD is responsive to teacher needs and more effective when teacher directed (Guskey and Yoon, 2009; Shulman, 1986).

The remaining 7 characteristics of effective professional development, identified by Thomas Guskey (2003) appeared on only one or two of the studied lists: effective PD… (15) helps accommodate diversity and promote equity (15%); (16) takes a variety of forms (15%); (17) provides opportunities for theoretical understanding (15%); (18) is driven by an image of effective teaching and learning (8%); (19) provides for different phases of change (8%); (20) promotes continuous inquiry and reflection (8%); and (21) involves families and other stakeholders (8%). Other research introduces other unique characteristics as well, such as, incorporating explicit theories of adult learning and the use of group settings into the delivery of PD (Elmore, 2002).

One such list detailing effective professional development was produced by Michael Puma and Jacqueline Raphael, in a *Evaluating Standards-Based Professional Development for Teachers: A Handbook for Practitioners*, prepared by The Urban Institute for the U.S. Department of Education (2001). The list summarizes components of effective professional development in detail and contains many of the characteristics listed above (2001, p. 10). "High Quality Professional Development:

- **Promotes an approach to teaching and learning that supports high standards for all students.** These approaches are aligned with standards and assessments. They can incorporate strategies for meeting the educational needs of diverse student populations. These strategies must be grounded in established knowledge about effective classroom teaching and learning and must be accessible to all educators.
- **Increases teachers' knowledge of specific content and of how students learn that content.** Deeping teachers' knowledge of specific disciplines that they teach is critical. Also important is the development of "pedagogical content knowledge" – professional development that focuses on the pedagogical implications of the discipline, such as understanding how students learn the discipline at different ages and in different contexts. Such professional development is rigorous and based on the knowledge base about teaching, as well as the underlying theory for that knowledge base.
- **Provides intensive, continuous, in-depth learning opportunities for teachers, with follow-up and support.** Professional development should include a high number of contact hours and span a long time period. These experiences should build on existing knowledge and permit teachers to collaborate, learn from each other and from external sources, experiment with new techniques, gain critical feedback, and continue to refine their teaching processes over a significant time period, in a continuous fashion, with repeated follow-up and support for ongoing learning as needed.
- **Expands the traditional role of teacher.** Current reforms demand that teachers take on new responsibilities to become leaders, mentors, peer coaches, curriculum and/or assessment designers, planners, and facilitators. In this environment of reform, teachers and other instructional staff form a community of learners who plan and work together to solve problems across the school and/or district. In addition, as many districts devolve authority to the school level, teachers are being asked to assume new roles in school governance and management (Corcoran, 1995). Teachers may be involved in identifying their professional development needs and in planning, designing, and delivering opportunities to meet those needs, as well as in assessing the effectiveness of these opportunities.
- **Connects directly to other reform programs and initiatives.** Professional development in the context of standards-based reform must be linked to other federal, state, district, and/or school initiatives. Such linkages can help to support teachers implementing new practices. The connection to school reform is also important to guarantee that professional development reflects specific local needs and abilities.
- **Is accountable for results.** Professional development should be evaluated regularly for its effect on teaching and learning. Multiple sources of data (e.g. teacher portfolios, classroom observations, peer evaluations, student performance) should be

used, with data collected at different times during the program implementation process. The results of these evaluations should be used to support continuous improvement.

- **Is collaborative.** By working together, teachers break down the isolation of individual classrooms and can begin to transform a whole school. Professional development activities should occur in groups of teachers from the same school, department, or grade level.
- **Is active and focused on problem solving.** Teachers need to be actively engaged in teaching and learning, particularly through curriculum development, action research, and other problem-solving activities."

Although there are many proposed criteria for effective professional development programs, it is important to acknowledge that there is no "one size fits all" model. One of the reasons that there are so many characteristics of effectiveness is because the fields of teaching and learning are highly complex endeavors embedded in equally complex systems (schools, districts, states, etc.). "These real-world contextual differences profoundly influence the effectiveness of professional development endeavors" (Guskey, 2003, p. 16). Each school and set of teachers is going to have unique needs and challenges that directly relate to selecting appropriate PD and assessing the quality of that PD. One school with a significant proportion of underprepared teachers may highly benefit from PD focused on classroom management and development of content knowledge, while another school with more experienced and accomplished teachers may see little benefit from the same program. The context in which the PD takes place is extremely influential and important in determining effective professional development (Elmore, 2002; Guskey and Yoon, 2009; Loucks-Horsley, Hewson, Love, and Stiles, 1998; National Staff Development Council, 2001).

Measuring Teacher Effectiveness

Just as effective teaching is defined in many ways, so too are the methods for measuring teacher quality and its effects. In general, teacher evaluation can be approached from three different angles: measurement of inputs; processes; or outputs (Goe, Bell, & Little, 2008). *Inputs* refer to that which the teacher brings to his or her position at the school, encompassing things like education, certification, knowledge, experience, etc. (Goe, Bell, & Little, 2008). The NCLB requirement for highly qualified teachers directly addresses measuring and improving teacher inputs. Processes refer to the interactions that take place between teachers and students or between the teacher and school or larger educational community (Goe, Bell, & Little, 2008). *Outputs* are the direct and indirect effects of *processes* and include impact on students, such as behavior, learning, and achievement, as well as, impacts on other teachers, or members of the school community (Goe, Bell, & Little, 2008). A broad conceptualization of teacher effectiveness refers to all facets that contribute to a teacher's success: inputs, processes, and outputs (Goe, Bell, & Little, 2008).

Traditionally, emphasis is placed on regulating and measuring teacher qualifications or inputs, with the assumption that adequate credentials and experience equals effective teaching. Research consistently shows that teachers matter and that effective teaching is dictated by much more than credentials and experience. Alternatively, emphasis is placed on assessing whether a teacher's performance is "satisfactory" or "unsatisfactory." Under this type of system, extremely high percentages of teachers receive "satisfactory" ratings, resulting in a poor job of differentiating teacher quality (Gordon et al., 2006). These typical evaluations are perfunctory compliance exercises that yield little useful information (The New Teacher Project, 2010). Stull evaluations (derived from the Stull Bill, AB 293), used in California, are an example of traditional teacher evaluation.

There are multiple flaws with traditional teacher evaluation systems. For the most part, evaluations are infrequent, especially for inexperienced teachers. Often teachers go consecutive years before receiving any meaningful feedback on performance (Weisberg et al., 2009). Even new teachers receive tenure after two or three years (18 months in California) in the teaching field with very little assessment of teaching practice or impact (Gordon et al., 2006). Evaluations can also be unfocused, looking at superficial behaviors and practices with little connection to student progress (Weisberg et al., 2009). Methods like the Stull performance evaluation are pass/fail systems, making it impossible to distinguish between levels of teaching.

What complicates this matter even worse is that nearly all teachers receive satisfactory ratings (Weisberg et al., 2009). This phenomenon has come to be known as the "Widget Effect – the tendency of school districts to assume classroom effectiveness is the same from teacher to teacher" (Weisberg et al., 2009, p. 4). Teachers often report that traditional evaluation systems are unhelpful because they provide no clear or useful feedback on performance in the classroom, nor are they tied to meaningful professional development opportunities (Weisberg et al., 2009). The last major flaw with typical teacher evaluation systems is that they are inconsequential. Rarely are teachers involuntarily discharged from schools based on poor performance, nor do the best teachers receive incentives to teach the highest need students (Gordon et al., 2006; Weisberg et al., 2009).

There are multiple measures that can be used to evaluate teacher effectiveness. Each measure is context-specific and appropriate under different conditions, garnering their own strengths and threats to validity. Generally speaking, measures gather evidence about growth in student learning and competency, evidence of instructional quality, and a range of other evidence based on local values (Goe, 2010). Some of the available measures include: content assessment; principal evaluation; master educator evaluation; peer evaluation; self evaluation; parent feedback; student feedback; surveys; logs; classroom observations; video assessment; teaching portfolios; student work; and student achievement. Some measures can be used formatively to improve teaching and learning (e.g. observation with follow-up) and help teachers grow (self-reflection), and some are too removed from classrooms to have much impact (value-added models) (Goe, 2010).

Increasingly, there is a push towards the "portfolio view" of teachers, where multiple measures of teacher effectiveness are used to determine teacher quality (Bausell, 2011; Gordon et al., 2006; Moss et al., 1998). Portfolio evaluation is an integrative assessment practice that pulls from the research tradition of hermeneutics, where information is combined across multiple pieces of evidence to interpret human products, expressions, or actions (Moss et al., 1998). This approach seeks to understand the whole in light of its parts, a holistic and integrative way to understand complex human phenomena (Moss et al., 1998). The suggested measures for inclusion in these models span from student achievement measures (most often value-added estimates, when available); personnel review (most often done by a Principal or other expert); self-evaluation; peer evaluation; student work; and student feedback. Portfolio evaluation is believed to be a system with inherently higher levels of validity because it takes a holistic approach to evaluating teachers (Goe, Bell, & Little, 2008). Portfolio evaluation is referred to here as a comprehensive teacher evaluation system (CTES).

There is currently much discussion over which measures to include in a CTES and how much weighting should be given to each measure. The Hamilton Project in collaboration with

The Brookings Institution recommend evaluating teachers using multiple measures of job performance, including: principal evaluation; parent evaluation; classroom observations; teacher attendance; and a measure of "value-added" (Gordon et al., 2006). Linda Darling-Hammond (2011) recommends a teacher evaluation system that both evaluates and supports teacher effectiveness. Her *Integrated Evaluation System* looks at teaching practice in relation to student needs, state standards, curriculum goals, contributions to colleagues and the school, and impact on learning. Dr. Darling-Hammond proposes using "standards-based observation; examination of curriculum plans, assignments, and student work samples; evidence of practices that support student learning (both inside and outside the classroom); and evidence of student learning measured in a variety of ways" (talk at CPECs Research Director's Meeting, 2011).

The National Board for Professional Teaching Standards is one organization that uses the portfolio method in evaluating teacher effectiveness. NBPTS portfolios include: "one classroom-based entry with accompanying student work; two classroom-based entries that require video recordings of interactions between [the teacher] and students; one documented accomplishments entry that provides evidence of accomplishments outside of the classroom and how the work impacts student learning" (2011, p.1). A four-hour content assessment is also required in the evaluation of effectiveness by the NBPTS.

A Framework for Teaching, a multi-faceted teacher evaluation protocol proposed by Charlotte Danielson (1996), was adopted by LAUSD as the basis for the evaluation of teacher effectiveness. Danielson's model covers four major domains of effective teaching: planning and preparation; the classroom environment; instruction; and professional responsibilities. Each standard contains multiple components, fully described in the documents published March 20th, 2011 (LAUSD). Shortly after, on March 24th, 2011, LAUSD published a 30-page rubric for evaluating teacher effectiveness based on the standards. For each of the components and elements, teachers may be rated as ineffective, developing, effective, or highly effective. The rating system is well aligned to the federal Race to the Top teacher evaluation system. The goal is to clearly define standards, components, and elements so that teachers and evaluators understand the new expectations and the resulting ratings. Paired with other measures of teacher effectiveness, the protocol can be used for formative (supportive of improvements) and summative (identifies the result or impact of teaching) evaluation purposes.

As we move towards an evaluation system that is both formative and summative it is wise to consider the "Keys to Measuring Teacher Effectiveness," in which, Laura Goe recommends measuring "what is required" and "what is valued," making known to teachers and evaluators the standards for evaluation and the "tools and processes of evaluation;" a system where performance assessments are well aligned with relevant standards (2010, p. 25). The New Teacher Project offers "six design standards that any teacher evaluation system must meet in order to be effective" (2010, p. 3). Similar to Goe, they recommend an annual process with clear and rigorous expectations involving multiple measures and multiple ratings connected to regular feedback and evaluation significance (The New Teacher Project, 2010). Recognizing the unique challenge faced by novice teachers, The New Teacher Project also recommends that expectations should steadily increase during a teacher's first few years in the classroom (2010).

Creating a teacher evaluation infrastructure that facilitates both an evaluative and supportive function benefits the field in multiple ways: decisions about tenure and pay will be easier; identifying areas for teacher development will be easier; identifying teachers that can serve as leaders and mentors will be easier; and the best approaches to teaching will be better identified (Gordon et al., 2006). Evaluation, done well, can increase professionalism and improve teaching, learning, and achievement (Lansman, 2006). I am in agreement with Laura Goe's belief that, "the ultimate goal of teacher evaluation should be to improve teaching and learning" (2010, p. 4). For an evaluation system to be fair and comprehensive it must be based on multiple measures of teacher effectiveness (Goe et al., 2008). This study further investigates three of the proposed measures for inclusion in a comprehensive teacher evaluation system.

CHAPTER 3: Methods

This study takes place in the context of an evaluation of a state-funded Improving Teacher Quality grant program (ITQP) in two Partnership for Los Angeles Schools (PLAS), in the Los Angeles Unified School District (LAUSD). The grant and corresponding evaluation spanned from October 2009 – September 2013. The evaluation of ITQP used a quasi-experimental design to assess the degree to which ITQP met its primary goals for teachers and students (around improving teacher effectiveness, establishing professional learning communities, and improving student achievement). A brief summary of the ITQ Program and evaluation is provided in Appendix A, to give further context to this investigation.

The purpose of this study is distinctly different from the evaluation of ITQP. This study is a methodological investigation of the measures used to gauge teacher effectiveness in relation to developing an estimate of overall teacher quality. Findings from this study directly inform the selection of teacher effectiveness measures for use in a comprehensive teacher evaluation system (CTES). Data for this study is pulled from a subset of the data from the ITQ Program, made available November 2013. This investigation takes place post-ITQP and is approved by the University of California, Los Angeles (UCLA) Institutional Review Board (IRB) (IRB#14-000154).

The ITQ Program context is a perfect setting for this studies investigation because significant involvement in ITQP APD is expected to increase teacher effectiveness, providing fertile ground for assessing each measures' sensitivity to assessing effectiveness, and capturing changes in effectiveness. This study assumes that the study teachers engaged in high quality professional development through the ITQ Program (this assumption is supported by the literature on effective professional development, ITQP program design and delivery, and study findings connecting APD with scores on teacher effectiveness constructs). This study does not assess the impact of ITQP but instead uses the context of the PD initiative to thoroughly study measures of teacher effectiveness.

This chapter begins with a description of the study sites, including school and community characteristics, student characteristics, and teacher characteristics. A brief summary of the study design and formation and description of the study sample, as well as instrumentation and data collection, follows. The chapter ends with an overall summary of analysis conducted, including specifics by research question and instrument. A complete list and description of independent and dependent variables used in the study is also provided.

Description of the Study Sites

This in-depth study of measures of teacher effectiveness involves science and social studies-history teachers at two middle schools in a large, urban school district (LAUSD). Each of these schools is a member of PLAS, otherwise known as the "mayor schools," which are a consortium of schools targeted by the mayor of Los Angeles for improvement. Both schools serve a diverse student population that includes significant populations of students from high-poverty backgrounds. What follows is a detailed description of the school, student, and teacher demographics. Data is presented for the three-years of study and one-year prior, for a baseline reading, pre-intervention.

School and Community Characteristics

The data used in this study is collected from teachers at two middle schools housed in the Partnership for Los Angeles Schools (PLAS) residing within the Los Angeles Unified School District (LAUSD): School A and School B. During the 2011-2012 academic year (program year 3) 3,904 students were enrolled in the schools A and B (Table 1). Both schools experienced a decline in total enrollment over four academic years. From 2008-2009 to 2011-2012, total enrollment decreased 34 and 15 percent at Schools A and B, respectively.

In academic year 2008-2009 (baseline year – pre-study commencement), both schools were ranked at the lowest score of 1 (high score is 10) for their statewide API (academic performance index) and neither school met its AYP (annual yearly progress) goal.

Table 1.
Enrollment in Schools A and B by Grade over Time

					Total	
School	School Year	Grade 6	Grade 7	Grade 8	Enrollment[a]	%Change
School A	2008-2009	466	575	581	1,622	
	2009-2010	451	467	529	1,449	-10.7
	2010-2011	332	440	443	1,215	-16.1
	2011-2012	307	353	412	1,072	-11.8
School B	2008-2009	704	762	816	2,282	
	2009-2010	673	749	697	2,120	-7.1
	2010-2011	545	745	749	2,040	-3.8
	2011-2012	552	674	719	1,945	-4.7
Total	2008-2009	1,170	1,337	1,397	3,904	
	2011-2012	859	1,027	1,131	3,017	-22.7

Note: Information for this table comes from DataQuest and CDE.
[a]Total enrollment includes "other grades"

Correspondingly, both schools were also in Year 5 of Program Improvement Status. In the final year (2011-2012) of the study, neither school achieved AYP. Both schools met API targets for one year out of the four examined. School A met targets in 2010-2011 and School B met API targets in 2011-2012. Schools A and B met API targets for all subgroups[2] for only one year of the four studied. Both schools were in Program Improvement status through most of the duration of the study, including 2011-2012 (Table 2).

Table 2.
AYP and API for Schools A and B by Academic Year

		Made AYP	Met API Target Schoolwide	Met API Target All Subgroups	PI status
School A					
	08-09				X
	09-10				
	10-11		X	X	X
	11-12				X
School B					
	08-09				X
	09-10	X			
	10-11	X			
	11-12		X	X	X

Note: This information was obtained from DataQuest and CDE.

Both schools are located in communities characterized by high crime rates, high poverty rates, and low levels of educational attainment (Table 3). In the period from November 2012 – May of 2013, the community around School A experienced an overall crime rate of 222.5 crimes per 10,000 people and School B experienced a total crime rate of 113 crimes per 10,000 people. According to the Los Angeles times, in both cases, these rates are higher than those of nearby communities (L.A. Times, 2013). In addition to higher instances of crime, these neighborhoods are also characterized by significantly lower educational attainment, lower median household income, and higher poverty rates than the city of Los Angeles as a whole.

[2] Subgroups include the following ethnic and socioeconomic categories: African American or Black (not of Hispanic origin), American Indian or Alaska Native, Asian, Filipino, Hispanic or Latino, Pacific Islander and White (not of Hispanic origin), plus socioeconomically disadvantaged. Students are categorized as socioeconomically disadvantaged if they participate in the federal free and reduced-price lunch program or if their parents did not graduate from high school.

Table 3.
Demographic Information for Communities Surrounding Schools A and B as Compared to the City of Los Angeles

	Surrounding Communities		
	School A[a]	School B[a]	Los Angeles[b]
Population			
Total population	25,797	99,243	3,792,621
People/ sq. mile	15,060	14,229	8,092
Ethnicities			
White (non-Hispanic)	1.0%	2.0%	29.4%
Black	39.3%	0.9%	9.6%
Latino	58.6%	94.0%	48.5%
Asian	0.4%	2.4%	11.3%
Other	0.7%	0.7%	1.2%
Foreign Born population			
Total	35%	52.4%	39.3%
Households			
Median household income	$29,897	$33,235	$50,028
Average household size	3.7	3.8	2.8
Households headed by a single parent	27.9%	21.9%	11.1%
Households below poverty line	26.0%[b]	31.0%[b]	20.0%
Educational Attainment			
Residents with a 4-year degree	3.9%	5.0%	20.2%
Residents with less than a high school degree	25.5%	33.9%	10.5%

[a]Los Angeles Times Mapping L.A. Project
[b]U.S. Census Bureau American Community Survey

Student Characteristics

The two middle schools serve a majority Latino student population that includes significant numbers of students from high-poverty backgrounds. During 2008-2009, 99% of School B students identified as Hispanic or Latino. For the same period, 69% of School A students identified as Hispanic or Latino, and 30% of students identified as Black or African American. Table 4 details student classification across school sites over time (LAUSD, 2012).

Each year, nearly all students qualify for Free and Reduced-Price Lunch (FRPL). There are never more than 10% of students included in the Gifted and Talented Education (GATE) program at either school site. Between 26% and 37% of students are identified as English Language Learners (ELL) across all four years. Between 8% and 18% of students are considered to have disabilities across all four years.

Table 4.
Student Demographics (Percentages) by Academic Year at Treatment and Comparison Schools

		Ethnicity		Student Classifications				
		Black or African American	Hispanic or Latino	FRPL[a]	GATE[b]	ELL[c]	RFEP[d]	SWD[e]
School A								
	08-09	30	69	87	4	33	24	11
	09-10	29	70	80	4	35	23	11
	10-11	32	68	99	4	29	25	13
	11-12	32	67	100	4	33	22	18
School B								
	08-09	0	97	87	10	37	43	8
	09-10	0	98	99	10	36	44	9
	10-11	0	98	99	9	31	47	10
	11-12	0	98	100	9	26	48	12

[a] Free and Reduced Price Lunch
[b] Gifted and Talented Education
[c] English Language Learner
[d] Reclassified Fluent English-Language Proficient
[e] Students with Disabilities

There is an observed achievement gap between ELLs and SWDs and the larger school populations at both schools. In 2008, while 34% of School A students scored at or above proficient on the 8th grade CST science exam, only 6% of students with disabilities and 13% of English learners earned comparable scores. This same trend was found at School B—when compared to school averages, fewer English learners and students with disabilities scored at or above proficient on the eighth grade science CST exam.

Similar trends in social studies-history CST scores were identified for each school. At School B, 19% of all students scored proficient or advanced on the history CST; however, only 5% of English learners and 1% of students with disabilities achieved similar scores. Furthermore, 2008 was the first time in the previous three years that a single SWD at School A scored above the basic level on the history CST. Finally, both school populations include many English learners and students with disabilities (in 2007-2008 these student groups accounted for approximately 40% of the total student population for both schools), yet neither school had met AYP goals for these student populations in the years preceding the study.

Teacher Characteristics

As Table 5 shows, most teachers at School's A and B have a Bachelor's degree. The number of teachers with a Master's degree gradually increased from 2008 to 2011, generally indicating, that the level of educational attainment among teachers at each school increased throughout the duration of the study. The number of teachers with a Doctorate degree is very small. Both schools experienced teacher attrition from 2008 to 2012.

Table 5.
Teachers' Education Level and Teaching Credentials

School Year	08-09	09-10	10-11	11-12	08-09	09-10	10-11	11-12
Education Level	School A				School B			
Total # of teachers	76	75	67	65	104	108	101	99
Doctorate	0	0	0	1	0	1	0	0
Mater's Degree +30	12	15	24	18	20	23	28	23
Mater's Degree	8	6	5	4	8	9	5	6
Bachelor's Degree +30	18	16	23	27	60	58	55	57
Bachelor's Degree	38	38	11	14	16	17	11	9
Less than Bachelor's Degree	0	0	1	0	0	0	0	1
Non Reported	0	0	3	1	0	0	2	3
Full Credential	69	NA	NA	NA	102	NA	NA	NA

Note: Information comes from DataQuest and CDE.

In addition to achieving higher levels of educational attainment, the teachers became more experienced through the duration of the study (Table 6). The average years of teaching experience increased in both schools from 2008 to 2012, comparable to district averages. Furthermore, the number of first- and second-year staff decreased during this period.

In addition to the identified disparity in CST exam scores, Schools A and B were targeted

Table 6.
Teachers' Average Years of Experience by Academic Year and School

School Year	08-09	09-10	10-11	11-12	08-09	09-10	10-11	11-12
Years of Experience	School A				School B			
Average Years of Service	5	6.1	7.9	9.9	11.7	12.2	12.4	13.8
Average Years in District	5	6.1	7.9	9.5	11.5	12	12.3	13.6
# First Year Staff	17	4	2	1	9	5	3	3
# Second Year Staff	24	20	5	4	12	7	4	3

Note: Information comes from DataQuest and CDE.

for improvement because both schools had high percentages of new teachers. Effective teachers must develop what Shulman (1986) considers two types of knowledge: 1) deep content knowledge of the subject itself and 2) pedagogical content knowledge. Over half of the teachers at School A and 25% of teachers at School B were classified as non-permanent; the majority having less than 5 years teaching experience. Because a large proportion of teachers were still at the early stages of their careers (at the beginning of the study), a need existed to develop these two types of knowledge. Teacher data along with the CST and AYP data showed a need for the focused training and support embodied in ITQP (Data obtained using LAUSD 07-08 School Report Cards & information on greatschools.com). Being that part of the focus of ITQP was to increase teacher effectiveness, this study is positioned to investigate the measures used to assess teacher effectiveness, including the extent to which each measure is sensitive to detecting change over time.

Study Design

This study seeks to create an understanding of what three distinct measures of teacher effectiveness, namely, a teacher survey, expert assessment, and classroom observations, can contribute to an understanding of teacher quality. This study is designed to conceptually and empirically explore the sensitivity of those measures to detecting differences (within-group and between-group) (RQ 1), relationships between teacher effectiveness constructs and measures (RQ 2), and the extent to which depictions of teachers vary across the different measures of effectiveness (RQ 3). Longitudinal data is pulled from a state funded Improving Teacher Quality (ITQ) professional development program for middle school science and social studies-history teachers to answer the research questions stated on p. 15.

Given the current landscape of teacher evaluation in education, understanding what each measure of teacher effectiveness contributes to an overall understanding of teacher quality is paramount. Articulations of teacher quality are directly impacted by the measures involved in ascertaining effectiveness making it imperative that measures are chosen and used appropriately. Only when we truly understand how to measure the multiple aspects of teacher effectiveness and how those measures in combination paint a picture of teacher quality can we safely use teacher evaluation findings for both formative and summative purposes. Findings are used to better understand measures for inclusion in comprehensive teacher evaluation systems, with the ultimate intent being to enhance the evaluation of teacher effectiveness and teacher quality.

Study Sample

The sample of teachers used in this study is a subset of the sample of teachers involved in ITQP. In the ITQP study there were 117 total teachers from schools A and B[3]. Because of the nature of the research questions in this study, analysis is restricted to mainstream science and

[3] Not including comparison school teachers – who were never considered for this study because limited data is available and teachers were not systematically tracked over time.

social studies teachers. With the elimination of Special Education Teachers[4] from the sample, 99 mainstream science and social studies teachers remain for consideration of inclusion in the study. Some of those teachers were only in ITQP for a short period of time and failed to participate in evaluation efforts, resulting in limited data for some teachers.

In order to be included in the sample for this study, teachers must have a minimum of two data points available for analysis. This could be two data points from the same instrument over time, or it could be two data points across instruments and/or time. Because this study is particularly interested in examining patterns across measures and/or across time, only teachers that would meaningfully contribute to such analyses are kept in the study[5]. With the elimination of teachers with only one data point available for analysis, 69 total teachers remain in the study.

Table 7, displays the number of teachers, broken down by school and content area, included in the study across three years. In any given year, there are between 59 (Year 1) and 66 (Year 2) total teachers actively participating in the study. School B has a higher number of teachers in the study, due to the higher enrollment at the school site and slightly lower teacher turnover than school A. Across both schools, study numbers remain relatively constant over time (between 19 to 24 teachers at School A and between 40 to 42 teachers at School B). Science and social studies teachers are equally represented in the study (between 30 to 33 science teachers and between 29 to 33 social studies teachers). With the high levels of teacher turnover at both schools, teachers remained in the study anywhere from one to three years. The total number of teachers in the study (69) exceeds the number of teachers by any given year, due to teacher turnover issues. Thus, as illustrated for School A, across the three years of the study

[4] Special Education teachers are not included in this study in order to minimize contextual factors that could cloud teacher effectiveness comparisons.

[5] If a teacher was only in the study during Year One, they are automatically excluded from the study because only one data source was collected at that time (Teacher Survey), eliminating the possibility of comparisons across measures.

(Y123 column) a total of 25 distinct teachers participated, which is higher than the total number of teachers in any one given year.

Table 7.
Number of Active Teachers in the Study across Years

	Year 1	Year 2	Year 3	Y123
School A				
Science	10	12	10	12
Social Studies	9	12	12	13
Total	19	24	22	25
School B				
Science	20	21	20	23
Social Studies	20	21	20	21
Total	40	42	40	44
Study				
Science	30	33	30	35
Social Studies	29	33	32	34
Total	59	66	62	69

Table 8 displays school site-specific teacher characteristics, pulled from the teacher survey used in this study. Grade representation remains relatively constant across years, with nearly half of the teachers teaching 6th grade and roughly a third of teachers teaching 7th grade and a third of teachers teaching 8th grade (teachers often teach multiple grades). In Year 1, nearly half of the teachers (47%) are new to the school site, with less than 3 years teaching experience at the school. School site establishment increases over time to 71% (Year 3) of teachers having taught at the school site for more than 3 years, indicating increases in teacher retention to some extent during ITQP.

Individual teacher characteristics are detailed in Table 9. Between 90% (Year 3) to 95% (Year 2) of teachers are classified as full-time teachers, the remaining 5% to 10% of teachers being part-time or long-term substitutes. The percentage of veteran teachers increases over time

Table 8.

School Site Specific Teacher Characteristics as a Percentage of the Study Sample over Time, Teacher Survey

	Y1 (n=45)	Y2 (n=64)	Y3 (n=61)
School			
A	36	38	34
B	64	63	66
Content Area			
Science	56	50	48
Social Studies	44	50	53
Grade[a]			
6th	47	47	44
7th	33	27	31
8th	36	30	28
Established at School Site			
New (<3 years)	47	31	29
Established (>3 years)	53	69	71

[a]Percentages total more than 100 percent because teachers often teach multiple grade levels.

from 71% in Year 1 to 95% in Year 3, as does the percentage of teachers possessing clear credentials, only 43% in Year 1 to 88% in Year 3. Corresponding to the increase in clear credentials, there is a decrease in teachers receiving Beginning Teacher Support and Assessment (BTSA) (from 57% in Year 1 to 12% in Year 3), the California new teacher induction program support. Both the increase in experienced and credentialed teachers over time points to an increasingly qualified (according to inputs) teaching staff at both schools. Nearly all teachers are qualified to teach the content area and grade level being taught (between 93% in Year 3 and 96% in Year 1). Very few teachers in the sample are National Board Certified (3% in Year 2).

The ITQ Program was offered to all science and social studies teachers at Schools A and B. PD activities included all-day PD (ADPD) and extended-day PD (EDPD), summer institutes, leadership building, peer coaching, and classroom observation. Given that the focus of PD was similar for ADPD and EDPD and that these two PD elements align with end of year data

Table 9.
Individual Teacher Characteristics as a Percentage of the Study Sample over Time, Teacher Survey

	Y1 (n=45)	Y2 (n=64)	Y3 (n=61)
Primary Position at School			
Full-Time Teacher	91	95	90
Part-Time Teacher	0	2	7
Long-Term Substitute	9	3	3
Teaching Experience			
Novice (<3 Years)	29	13	5
Veteran (>3 Years)	71	87	95
Credential Status			
No Credential	0	2	2
Preliminary Credential	57	20	10
Clear Credential	43	79	88
Credential Type[a]			
Multi-Subject Matter	48	50	41
Single-Subject Matter	32	42	47
Qualified to Teach Content and Grade			
No	4	5	7
Yes	96	95	93
BTSA Support			
No	43	79	88
Yes	57	21	12
National Board Certified			
No	-	97	100
Yes	-	3	0

[a] Percentages do not total to 100 because teachers may or may not have one or more credentials.

collection, these two program elements are often referred to as academic year PD (APD). Attendance was taken at all academic year and summer PD sessions. Since participation was voluntary, not all teachers elected to participate in the program. Table 10 shows the minimum, maximum, and average participation in ITQP by school and content area over the life of the program. Numbers include the PD offered during the academic year and the summer professional development.

Table 10.
Total Range and Average Hours of Participation in PD for Study Teachers

		Academic Year PD				Summer Session PD				All PD			
	n	Min	Max	M	SD	Min	Max	M	SD	Min	Max	M	SD
School A													
Science	12	0	120.0	47	(39.13)	0	33.0	6	(11.51)	0	150.5	53	(48.50)
Social Studies	13	0	99.0	35	(31.85)	0	49.0	14	(17.11)	0	127.0	49	(43.62)
Total	25	0	120.0	41	(35.27)	0	49.0	10	(14.92)	0	150.5	51	(45.09)
School B													
Science	23	0	118.5	42	(43.22)	0	62.0	11	(22.31)	0	180.5	53	(58.33)
Social Studies	21	0	85.5	44	(21.97)	0	47.0	8	(13.26)	0	116.0	53	(29.88)
Total	44	0	118.5	43	(34.38)	0	62.0	10	(18.41)	0	180.5	53	(46.44)
Study													
Science	35	0	120.0	43	(41.35)	0	62.0	10	(19.28)	0	180.5	53	(54.43)
Social Studies	34	0	99.0	41	(26.13)	0	49.0	10	(14.83)	0	127.0	52	(35.17)
Total	69	0	120.0	42	(34.47)	0	62.0	10	(17.11)	0	180.5	52	(45.64)

Note. Academic Year PD is comprised of All-Day PD (ADPD) and Extended-Day PD (EDPD). Each ADPD session was 7 hours in length. Each EDPD session was 1.5 hours in length.

Overall, there is great variation in cumulative levels of ITQP participation. Maximum involvement over the life of the program for Academic PD is 120 hours, for Summer PD, 62 hours, and for all PD combined, 180.5 hours. Some teachers did not attend any PD sessions over the life of the program. M, mean number of participation hours are similar between schools (41 average APD hours for School A and 43 APD hours for School B), and at both schools, science teachers generally participated in more hours than did social studies teachers (120 maximum APD hours versus 99 maximum APD hours for social studies). Knowing the level of

PD participation is essential to answering the research questions for this study (APD is used as a continuous predictor variable and facilitates comparisons across participation groups).
Based on the ITQP program theory regarding intensity of APD participation, three groups are formed: no/low participation (0 - 24.5 APD hours), moderate participation (25 - 73.5 APD hours), and high participation (74 -120 APD hours). For teachers in the no/low participation group, the ITQ Program does not expect to see any intended PD effect. As the strength of treatment increases so too does the expected impact. For teachers in the high participation group the ITQ Program expects to see the greatest level of effect.

Teachers in the moderate and high participation groups were eligible for stipends (ranging from $75.00 to $1,000, yearly). In order to receive a minimum stipend, a teacher needed to attend half of the EDPD sessions (7), and at least 2 ADPD sessions, within a year. As APD participation increased so too did the eligibility for a greater stipend. Table 11 shows the participation characteristics of teachers belonging to each of the participation groups. Average hours of APD participation are: 8 for the no/low group; 50 for the moderate group; and 100 for the high group.

Table 11.
Academic Year PD Participation Characteristics of Low, Moderate, and High Participation Groups

	n	Min	Max	M	SD
No/Low Participation	27	0	22.5	8	(8.55)
Moderate Participation	30	29	69	50	(10.63)
High Participation	12	78.5	120	100	(15.78)
Study	69	0	120	42	(34.47)

Note. Extent of Academic Year PD (APD) participation over the life of the program is taken into account to form three groups: no/low participation (0-24.5 APD hours), moderate participation (25-73.5 APD hours), and high participation (74-120 APD hours).

Instrumentation and Data Collection

In total, three data sources are used to investigate sensitivity, relationships, and depictions of teachers across measures of teacher effectiveness: a Teacher Survey, an Expert Assessment, and a Classroom Observation Protocol. Table 12 shows the timeline of data collection by data source as part of the evaluation of ITQP. All data was made available in November 2013, at the end of reporting for ITQP. Each data source and the methods used for collecting the data are described in greater detail below. To view the actual instruments, please turn to the appendices (Appendix B: Teacher Survey, p.174; Appendix C: Construct Reliability Coefficients and Item Factor Loadings for Teacher Survey, p. 181; Appendix D: Expert Assessment, p. 182; Appendix E: Classroom Observation – Standards, Components, and Elements, p. 183).

Table 12.
Outcome Measures by Data Collection Timeline

	Year 1	Year 2	Year 3-Time 1	Year 3-Time 2
Teacher Survey	Spring 2010	Spring 2011	-	Spring 2012
Expert Assessment	-	Spring 2011	Fall 2011	Spring 2012
Classroom Observation	-	-	Fall 2011	Spring 2012

Teacher Survey

The teacher survey (Appendix B) used in this study was developed and refined by the UCLA SRM Evaluation Group on several projects designed to evaluate the effects of professional development on teachers. Informed by multiple survey sources (Quartz, 2007; Mintrop, Heinrich, & Trujillo, 2007; Tourkin et al., 2007), evaluators, educational researchers, PD program directors, expert subject matter professionals, professional learning partners, and teachers, the instrument covers a broad base of teacher experiences and outcomes. Results from a confirmatory factor analysis (CFA) demonstrate that the survey measures several constructs (Instructional Efficacy, Collegiality, Leadership, Collaboration, Reflective Practice, and Commitment to Ongoing Learning) related to effective teaching. Construct reliability coefficients and factor loadings are presented in Appendix C. The survey spans seven-single-sided pages and in addition to the above constructs, contains demographic, monitoring, and satisfaction items. Table 13 provides examples of constructs and items contained in the survey.

Table 13.
Examples of Attitudinal and Behavioral Constructs and Items Contained in the Teacher Survey

Instructional Efficacy
In general, my classes are well behaved.
Students know that I expect hard work from them and they act accordingly.
Collegiality
There is a great deal of cooperative effort among staff members here.
I can count on colleagues here when I feel discouraged about my teaching or students.
Leadership
I am involved in decision-making for my school.
I have organized and presented professional development at my school.
Collaboration
I have collaborated with teachers within my department and my grade level.
I have taught a common lesson created collaboratively within my department.
Ongoing Learning - Conferences
I have attended professional conferences.
I have been the presenter at professional conferences.
Reflective Practice – Video Taping
I have videotaped myself teaching.
I have reviewed and reflected on a videotape of my teaching.

The Teacher Survey was administered in paper format at three time points throughout ITQP (Spring 2010, 2011, and 2012). Researchers went to each campus during school-wide or department meetings, or after school hours to administer the survey in person to science and social studies teachers. Teachers were offered incentives to participate in the survey, including: gift cards to a local coffee shop, entry into an opportunity drawing for an Amazon gift card, school supplies, and a pizza party (incentives provided to teachers had a monetary value of approximately $5 per teacher, per survey administration).

Expert Assessment

The Expert Assessment (Appendix D) was developed by the UCLA SRM Evaluation Group in consultation with the UCLA Center X PLPs. The Expert Assessment is designed to assess the frequency with which teachers are implementing various sets of instructional strategies on a scale from 1-4. The assessment tool divides instructional strategies into several categories correlated with effective teaching practices, specifically: reading, writing, inquiry, collaboration, and other strategies.

PLPs were instrumental in both the development of the Master Teacher Assessment, and with data collection. Researchers convened a daylong meeting with the PLPs to develop the instrument. Drawing on over 15 years of UCLA Center X's professional development work impacting student learning through increasing teachers' knowledge and use of instructional strategies, PLPs brainstormed an exhaustive list of strategies, and helped categorize them according to the distinctions above. During this meeting, PLPs and researchers also came to a consensus about use of the scale to describe the frequency with which strategies were employed. Table 14 provides examples of instructional strategies, organized by classification type.

Table 14.
Examples of Classification and Types of Instructional Strategies Guiding Expert Assessment

- Reading Strategies
  - Read-aloud
  - Paraphrasing
  - Say something
- Writing Strategies
  - Graphic organizers
  - Cornell notes
  - Sentence starters
- Inquiry Strategies
  - Analyzing documents
  - Discrepant events
  - Point of view
- Collaborative Strategies
  - Think-pair-share
  - Debates
  - Peer assessment
- Other Strategies
  - Project-based learning
  - Presentations
  - Realia

The PLPs for each subject area completed the instrument for their respective teachers. Given that the PLPs are science and social studies teaching experts, and have spent a significant amount of time with the teachers in the study, they are uniquely positioned to give a grounded and expert assessment of the teaching practice that they observe. This instrument was completed at three time points in the ITQ program: Spring 2011, Fall 2011, and Spring 2012.

Classroom Observation Protocol

The Classroom Observation Protocol consists of a variety of items that describe the classroom environment and delivery of instruction (Appendix D). This tool was informed by LAUSD's Teaching and Learning (T&L) Framework (Los Angeles Unified School District, 2011), modified from Charlotte Danielson's work around developing a Framework for Teaching (Danielson, 2007). The T&L Framework consists of five standards—planning, classroom environment, instructional delivery, professional responsibilities, and professional growth. Each standard contains nested components and elements, which further articulate standards of teaching practice.

The Classroom Observational Protocol used in this study is based on the two observable standards from the T&L Framework—classroom environment and instructional delivery. Under these two standards, there are a total of nine distinct components and thirty-one elements. As recommended in the federally funded Race to the Top (RTTT) teacher evaluation system, the T&L Framework Rubric uses a four-point scale with clear definitions of *ineffective*, *developing*, *effective*, and *highly effective* teaching for each observed element. Table 15 lists the standards and components used in this study.

Classroom observation data was collected in the third year of the program at one or two

Table 15.
Classroom Observation Standards and Components

(S2) Classroom Environment	
	(S2CA) Creating an environment of respect and rapport
	(S2CB) Establishing a culture for learning
	(S2CC) Managing classroom procedures
	(S2CD) Managing student behavior
(S3) Instruction	
	(S3CA) Communicating with students
	(S3CB) Using questioning and discussion techniques
	(S3CC) Structures to engage students in learning
	(S3CD) Delivery of instruction
	(S3CE) Demonstrating flexibility and responsiveness

time points (Fall 2011 and/or Spring 2012) for participating teachers. The two primary ITQP researchers, both having received 40 hours of training on proper use of the observation protocol (and having obtained LAUSD Observer Certification) collected the data. Researchers went in to the classroom with laptop computers and objectively documented the activities and surroundings during each 55-minute class period. After leaving the classroom, the researchers entered ratings on an online version of the rubric. Additionally, the researchers categorized and counted the instructional strategies (based on the distinctions described above) that they observed teachers employing during the observation period.

All science and social studies teachers at the schools A and B were asked to participate in the classroom observations regardless of their attendance at ITQ PD. The researchers recruited participants in two ways: first, researchers attended APD sessions in order to establish rapport with teachers and build a trusting relationship that would facilitate researcher presence in the classroom. Second, researchers attended social studies and science department meetings to reach teachers that were not participating in the grant-sponsored PD.

Study participation was also incentivized in a variety of ways. First, each teacher was offered a transcript from the observation. This was particularly helpful for teachers that combined the observation with their filming for TakeOne! (which requires that a transcript be submitted with the video). During the first round of observations (Fall, 2011), each teacher that participated in an observation had their name entered into an opportunity drawing for a $100 gift card. One gift certificate per subject area per school was given out (totaling 4 prizes). Given low numbers of participation in the fall, the incentive structure was changed such that each participating teacher was entered into an opportunity drawing for an iPad 3 in the Spring (only one iPad was given out).

Analysis

Data are analyzed by instrument and across instruments to answer research questions. Analysis involves exploring associations between variables, looking for change over time, and testing for group differences between low, moderate, and high participation groups (RQ1). Relevant significance testing is conducted as well as statistical adjustments for running multiple tests of significance. Where the relationship between variables is strong, predictive power is also explored (RQ2). Pattern analysis is conducted within and across measures to identify trends in quintile placement for teachers and to assess comparability of measures (RQ3). Each analysis is explained in full detail in the results section (and summarized below). Findings are synthesized by teacher effectiveness measure and research question.

Specific analysis for each research question and instrument is explained below, in Table 16. A list and description of the independent and dependent variables of interest for this study can be found in Table 17 and Table 18. When possible, the most sophisticated analyses are conducted (e.g. between and within subjects design), but in some instances, the data does not lend itself to these procedures. In all cases, the most rigorous statistical procedure is selected.

For each procedure, all necessary assumptions are tested and problems addressed, if any (complete results from testing of assumptions are included in Appendices F and G). Additional details of analyses are presented in the findings section. Conclusions across instruments are presented in both the findings and discussion section.

Table 16.

Summary of Research Questions and Analysis by Instrument used for this Study

Research Question		Instrument	Analysis
1	What teacher effectiveness measure(s) exhibit(s) the greatest level(s) of sensitivity to detecting differences?	All	Summary of findings across 1a and 1b analyses.
1a	How does each measure capture changes in teacher effectiveness over time?	Teacher Survey	Repeated Measures ANOVA for the low, moderate, and high participation groups to test for significant change over time (T1, T2, T3) on instructional efficacy, collegiality, leadership, ongoing learning, and collaboration. Post hoc analysis with a Bonferroni adjustment when appropriate. Additionally, a paired samples t-test to test for significant change over time on reflective practice scores (T2 to T3). Measures of effect size are also calculated.
		Expert Assessment	Mixed ANOVA to determine whether there are differences between groups over time on their overall strategy use. Exploration of the simple main effects for time using 3 repeated measures ANOVAs and Huynh-Feldt statistics. Pairwise comparisons to detect significant growth over time for each group, with a Bonferroni adjustment. Measures of effect size are also calculated.
		Classroom Observation	For the 16 teachers with 2 observations, paired sample t-tests are used to test for significant growth over time on classroom environment, instruction, and total strategy usage (and corresponding sub-constructs) for each participation group (low, moderate, high).
1b	How does each measure capture group differences between low, moderate, and high participation teachers?	Teacher Survey	One-way MANOVA between low, moderate, and high participation groups is conducted to test for a significant difference in the vector of the means between groups at T1, T2, and T3. At T3, follow-up univariate ANOVAs with a Bonferroni correction indicate constructs where group differences exist and Games-Howell post-hoc tests indicate for which groups there are significant differences in construct scores. Measures of effect size are also calculated.
		Expert Assessment	Mixed ANOVA to determine whether there are differences between groups over time on their overall strategy use. Exploration of the simple main effects for group using 3 one-way ANOVAs. Follow-up univariate analysis of variance (Tukey HSD) to test for group differences in overall strategy use at T1, T2, T3. Measures of effect size are also calculated.
		Classroom Observation	T1/T2 averages are used when possible to create average scores on constructs and sub-constructs. MANOVA is used to determine if there are group differences on total strategy usage. One-way ANOVAs are used to determine if there are group differences on classroom environment and instruction.

Table 16.
Continued

Research Question		Instrument	Analysis
2	What are the relationships between measures of teacher effectiveness and teacher characteristics?	All	Summary of findings across 2a, 2b, and 2c analyses.
2a	What is the strength and direction of association between variables[a]?	All	A Pearson correlation table is created using the 10 continuous teacher effectiveness constructs and APD. Close inspection reveals the strength and direction of association between the dependent variables and APD used in the study. Regression analysis reveals the relationships between the independent variables (teacher characteristics) and dependent variables used in the study. Regression coefficients are inspected, revealing both the direction of association and strength.
2b/ 2c	In what ways are teacher characteristics predictive of teacher effectiveness? In what ways are academic PD participation and/or total strategy use predictive of teacher effectiveness?	All	A series of hierarchical multiple regression analyses are run (instructional efficacy, collegiality, leadership, reflective practice, overall strategy use, classroom environment, instruction), to control for the effects of the covariates (teacher characteristics) and determine the predictive power of APD or total strategy use. Multiple regression is used to test the predictive power of teacher characteristics on ongoing learning and collaboration. Year 3 factor scores generated by confirmatory factor analysis are used for the 6 teacher survey effectiveness constructs.
3	In what ways are individual teachers depicted similarly and differently across different measures of teacher effectiveness?	All	Standard scores are used to facilitate comparisons across constructs and measures. Inspection of quintile placement on 21 teacher effectiveness constructs and 3 composite constructs is used to discern patterns within and across teachers. Teachers ranking in the 10th and 90th percentiles within constructs are also noted and studied to discern patterns. Patterns are described both quantitatively (percent of the time above (quintile 4/5) or below (quintile 1/2) the mean) and qualitatively. Summary of findings from 3a and 3b.
3a	In what ways do these patterns hold across teacher effectiveness constructs?	All	Inspection of quintile placement for each case across constructs housed within an individual measure is used to understand patterns. 3 patterns analyses are conducted: quintile placement on classroom observation constructs and composite constructs (using average classroom observation scores); quintile placement on expert assessment constructs and composite construct at Spring Y2 (T1) and Spring Y3 (T3); and quintile placement on teacher survey constructs at Spring Y1 (T1), Spring Y2 (T2), and Spring Y3 (T3).
3b	In what ways do these patterns hold across measures?	All	Inspection of quintile placement for each case across constructs and measures is used to understand patterns. 2 pattern analyses are conducted: quintile placement in Y2 on the teacher survey and expert assessment constructs; and quintile placement in Y3 on the teacher survey, expert assessment, and classroom observation constructs.

[a] Table 17 lists and defines the independent variables included in this study. Table 18 lists and defines the dependent variables included in this study. Relationships between dependent variables and between independent and dependent variables are explored to the greatest extent possible in this study.

Table 17.

Independent Variables Included in this Study, Teacher Survey

Independent Variable	Response	Description	Category	Hypothesis
Credential Status	None, Preliminary, Clear	Credential status during a given year.	Input	Research suggests that the more qualified the teacher, the greater the effectiveness. Therefore, I expect a National Board Certified, Clear Credentialed Teacher with the proper qualifications to teach the content area and grade level being taught to have the highest level of effectiveness across measures.
Qualified to Teach (Grade and Content)	No, Yes	6th grade teachers should have a Multi-Subject Matter Credential. 7th and 8th grade teachers should have a Single Subject Matter Credential.	Input	
National Board Certified	No, Yes	National Board Certification Status.	Input	
BTSA Support	No, Yes	Beginning Teacher Support and Assessment (BTSA) is a New Teacher Induction Program.	Intervention	Research suggests that the more support teachers receive, the better they perform. I expect that with higher levels of participation in either intervention, we will see higher levels of effectiveness across measures.
Total APD	0-120	Total Academic PD (APD) is a continuous variable that includes sum total ADPD and EDPD participation hours over three years.	Intervention	
School	A, B	School A or B.	Context	Research suggests that contextual factors may impact teacher effectiveness. I expect School B teachers to outperform School A teachers because of the lower level of school chaos. I do not expect content area to be a significant predictor of effectiveness but include it for potential impact. I expect greater effectiveness from Full-time teachers because they experience the least amount of job uncertainty and highest benefits.
Content	Science, Social Studies	Content area science or social studies.	Context	
Position	Full-Time, Part-Time, Long-Term Substitute	Primary position at the school site during a given year.	Context	
Veteran Teacher	No, Yes	A veteran teacher has more than 3 years full-time teaching experience. A novice teacher has less than 3 years full-time teaching experience.	Experience	Research suggests that the greater the teaching experience, the greater the effectiveness. I expect teachers that are well established in the teaching field and at their school sites to have the highest levels of effectiveness across measures.
Established Teacher	No, Yes	An established teacher has taught full-time at their school site for over 3 years. A newly established teacher has taught full-time at their school site for less than 3 years.	Experience	

Table 18.

Dependent Variables Included in this Study, Teacher Survey, Expert Assessment, and Classroom Observations

Dependent Variable/ Construct	Item Response	Description and Relevance	Category	Data Source
Instructional Efficacy	1-5: Strongly Disagree, Moderately Disagree, Neutral, Moderately Agree, Strongly Agree	A teacher's belief in their ability to teach well. This includes organizing and executing lessons with competence to reach desired student outcomes (Hoy & Woolfolk, 1993). McLaughlin & Marsh (1978) point out how a teacher's belief in their own teaching ability is consistently related to student achievement.	Attitudinal	Teacher Survey
Collegiality	1-5: Strongly Disagree, Moderately Disagree, Neutral, Moderately Agree, Strongly Agree	According to Bang-Jensen (1986), "One of the most effective methods of providing that system (one through which teachers are revitalized, encouraged, and challenged) seems to be through a collegial support group. Within the confines of such a group, teachers could begin to think of one another as resources." (As cited in Gilman, Emhuff, & Hamm, 1998, p. 1).	Attitudinal	Teacher Survey
Leadership	1-8: Never, Yearly, Semesterly, Bi-Monthly, Monthly, Bi-Weekly, Weekly, Daily	"Teacher leadership is the process by which teachers, individually or collectively, influence their colleagues, principals, and other members of the school community to improve teaching and learning practices" (Yorr-Barr & Duke, 2004, p. 287). Harris & Muijis (2004), in their exploration of teacher leadership and school and classroom improvement, conclude that the role teachers play in spearheading improvement both in the classroom and school is critical in improving student achievement.	Behavioral	Teacher Survey
Ongoing Learning	1-7: 0 times, 1 time, 2 times, 3 times, 4 times, 5 times, 6 times or more	A continual pursuit of knowledge to improve teaching. In this study, ongoing learning is assessed through a teacher's involvement in national educational conferences believed to improve teaching and learning.	Behavioral	Teacher Survey
Reflective Practice	1-8: Never, Yearly, Semesterly, Bi-Monthly, Monthly, Bi-Weekly, Weekly, Daily	Schon (1993), introduces the idea of reflective practice as, "the capacity to reflect on action so as to engage in a process of continuous learning". This professional self-reflection leads to insight and change to improve teaching and learning. In this study, this is assessed through teachers' videotaping practices.	Behavioral	Teacher Survey
Collaboration	1-8: Never, Yearly, Semesterly, Bi-Monthly, Monthly, Bi-Weekly, Weekly, Daily	The process in which teachers work together to improve teaching by making connections across curriculum. Collaboration between teachers is a contributing causal factor in improving teachers' knowledge, skills and practices (DuFour, 2004).	Behavioral	Teacher Survey

Table 18.

Continued

Dependent Variable/ Construct	Item Response	Description and Relevance	Category	Data Source
Overall Strategy Use	1-4: Barely to Never, Sometimes, Often, Most of the Time to Always	Effective teachers use many available resources and call on multiple methods to teach their students and meet their goals. (Goe, et al., 2008; National Board for Professional Teaching Standards, 1989). This includes, among other things, having a repertoire of teaching strategies (Darling Hammond, 2011; Kemp & Hall, 1992).	Behavioral	Expert Assessment
Total Strategy Use	Count of Distinct Strategies		Behavioral	Classroom Observation
Classroom Environment	1-4: Ineffective, Developing, Effective, Highly Effective	As Charlotte Danielson points out, "The classroom environment is a critical aspect of a teacher's skill in promoting learning" (2007, p. 64). Effective teachers have control over their classrooms, establish collaborative norms, and structure environments to make learning possible (Kemp & Hall, 1992; Taylor, et al., 1999)	Behavioral	Classroom Observation
Instruction	1-4: Ineffective, Developing, Effective, Highly Effective	Instruction is at the heart of teaching. It is where an effective teacher engages students in learning, makes content accessible, and adapts teaching to student needs (Danielson, 2007; Darling-Hammond, 2011; Kemp & Hall, 1992; Taylor, et al., 1999).	Behavioral	Classroom Observation

CHAPTER 4: Results

This study seeks to create an understanding of what three distinct measures of teacher effectiveness, namely, a teacher survey, expert assessment, and classroom observations, can contribute to an understanding of teacher quality. Relevant findings, produced from investigating: the sensitivity level of each instrument (to detecting both between-group and within-group differences) (RQ 1); the relationships between teacher effectiveness constructs (measures, and teacher characteristics) (RQ 2); and the extent to which depictions of teachers vary across measures (RQ3), can be used to inform the design and implementation of comprehensive teacher evaluation systems (CTESs).

This chapter begins by looking at the data sources available for each teacher in the study and the implications for possible analyses. Findings follow, organized by research question and instrument. Each section starts with a review of the research question and sub-questions to be addressed and an overview of the subsequent analyses used. At the end of the findings presentation for each research question, a summary across measures is included. At the beginning of Chapter 5, a summary of results across research questions is provided. For all statistical analyses and subsequent tables and figures the following notation is used: * = $p< .05$; ** = $p< .01$; *** = $p< .001$. (In cases where a Bonferroni correction is used, p values are adjusted based on the number of statistical tests performed.)

Available Data Sources

Table 19 and Table 20 show the 69 teachers (by school and content area) in the sample and the data sources available for use in this investigation. For all 69 teachers in the study an investigation of significant change over time is available for one or more measures. Comparisons are made between participation (dosage) groups (teachers with low/no, moderate, and high levels of APD participation), and within participation groups to detect each measures' sensitivity to differences. Relationships between variables included in the study are also fully explored. All 69 teachers included in the study are included in at least 1 single measure pattern analysis. For the 54 teachers that have a minimum of two different data sources available, some form of pattern analysis across measures is possible.

Table 19.
Sample of Teachers at School A by Content Area across Measures and Time

	Teacher Survey			Expert Assessment			Classroom Obs.	
Teacher	Y1T2	Y2T2	Y3T2	Y2T2	Y3T1	Y3T2	Y3T1	Y3T2
Science								
1	-	1	1	-	-	-	-	-
2	1	1	1	-	-	-	-	1
3	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	-	-
5	1	1	1	-	-	-	-	-
6	1	1	1	1	1	1	1	1
7	1	1	1	-	-	-	-	-
8	1	1	-	-	-	-	-	-
9	1	1	-	-	-	-	-	-
10	1	1	-	-	-	-	-	-
11	1	1	1	1	1	1	-	1
12	-	1	1	1	1	1	-	-
Social Studies								
13	-	1	1	1	1	1	-	-
14	-	1	1	1	1	1	-	1
15	-	1	1	1	1	1	-	-
16	-	-	1	-	1	1	-	1
17	-	1	1	1	1	1	-	-
18	1	1	1	1	1	1	1	1
19	-	1	1	1	1	1	-	1
20	1	1	1	1	1	1	-	1
21	1	1	1	1	1	1	-	-
22	1	1	-	-	-	-	-	-
23	1	1	-	1	1	1	-	-
24	1	1	1	1	1	1	1	1
25	1	1	1	1	1	1	-	-

Note. Missing data is represented by a dash, symbolizing unavailability due to both unwillingness to participate and inactivity due to withdrawal from the sample.

Table 20.

Sample of Teachers at School B by Content Area across Measures and Time

	Teacher Survey			Expert Assessment			Classroom Obs.	
Teacher	Y1T2	Y2T2	Y3T2	Y2T2	Y3T1	Y3T2	Y3T1	Y3T2
Science								
26	-	1	1	1	1	1	-	-
27	-	-	1	1	1	1	1	-
28	-	-	1	1	1	1	-	-
29	-	-	1	1	1	1	1	-
30	1	1	-	-	-	-	-	-
31	-	1	1	1	1	1	-	-
32	1	1	1	-	-	-	-	-
33	1	-	1	-	-	-	-	1
34	1	1	1	1	1	1	-	-
35	1	1	1	1	1	1	-	-
36	1	1	1	1	1	1	1	1
37	1	1	1	1	1	1	-	-
38	1	1	1	1	1	1	-	-
39	1	1	-	-	-	-	-	-
40	1	1	1	1	1	1	-	1
41	-	1	1	-	-	-	-	-
42	-	1	1	1	1	1	-	-
43	1	1	1	1	1	1	-	-
44	1	1	1	1	1	1	1	1
45	1	1	1	1	1	1	1	1
46	1	1	1	1	1	1	-	1
47	1	1	1	1	1	1	1	1
48	1	1	-	-	-	-	-	-
Social Studies								
49	1	-	1	-	-	-	-	-
50	-	1	1	1	1	1	-	-
51	1	1	1	1	1	1	1	-
52	1	-	1	1	1	1	1	1
53	-	1	1	1	1	1	1	1
54	1	1	1	1	1	1	-	1
55	1	1	1	1	1	1	1	1
56	-	1	1	-	-	-	-	-
57	-	1	1	1	1	1	-	-
58	1	1	-	1	1	1	-	-
59	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	-	1
61	1	1	1	1	1	1	1	1
62	1	1	1	1	1	1	-	1
63	-	1	1	1	1	1	-	-
64	1	1	1	1	1	1	1	1
65	1	1	1	1	1	1	-	-
66	1	1	-	-	-	-	-	-
67	1	1	1	1	1	1	-	-
68	-	1	1	1	1	1	1	1
69	-	1	1	-	-	-	-	-

Note. Missing data is represented by a dash, symbolizing unavailability due to both unwillingness to participate and inactivity due to withdrawal from the sample.

Research Question 1

The following research question and sub-questions are addressed in this section: (1.) What teacher effectiveness measure(s) exhibit the greatest level(s) of sensitivity to detecting differences? (a.) How does each measure capture changes in teacher effectiveness over time? (b.) How does each measure capture group differences between low, moderate, and high (ITQP APD) participation teachers?

An investigation into the sensitivity level of each measure of teacher effectiveness requires looking first at data at the instrument level and then making comparisons across instruments. Specifically, I am looking at an instrument's sensitivity to detecting between-and-within-group differences over time. Wherever possible, both participation group (low, moderate, and high) (between-subject effect) and time (T1, T2 or T1, T2, T3) (within-subject effect) is used in the same analysis, minimizing the error introduced by running multiple individual tests of significance. Dependent variables of interest vary by instrument (see Table 18, above). Prior to each analysis, relevant assumptions are tested – the results are reported in Appendix F.

Teacher Survey

There are three time points (Y1T2, Y2T2, and Y3T2) available for analysis of the teacher survey data. Unfortunately, due to high levels of teacher turnover between year's 1 and 2 of the program and incomplete data for all cases, a mixed ANOVA analysis, which could look at between-group and within-group differences over time on teacher effectiveness constructs, is not possible[6]. Separate analyses are conducted to test for participation group differences on

[6] In order to proceed with a series of mixed ANOVA's, half of the sample (49%) would have been excluded from the analyses. Any benefit gained from the more rigorous statistical test would have been lost in inferior sample representativeness.

construct scores and within-group change over time. A one-way MANOVA between low, moderate, and high participation groups is conducted (including post-hoc analyses where appropriate) to test for differences in the vector of the means between participation groups at T1, T2, and T3. A MANOVA test is useful because it combines the dependent variables (5-6 constructs) to form a new linear composite variable that maximizes the differences between groups. Additionally, repeated measures ANOVAs (including post-hoc tests when appropriate) are used to test for significant change over time (T1, T2, T3) for each of the participation groups on teacher effectiveness constructs. Paired samples t-tests are used for all three groups to test for significant change over time in reflective practice scores (T2, T3). The overarching expectation is that higher participation groups will have higher scores on teacher effectiveness constructs and that those scores will increase at greater rates (than lower participation groups) over time.

Testing for Within-Group Differences on the Teacher Survey

Table 21 displays the mean construct scores for the three treatment groups over time. For each of the three groups, there are mixed trends in scores across years. Interestingly, all three participation groups see a decline in ongoing learning mean scores over time. For the no/low participation group, scores are generally similar across the three years with the exception of collaboration (which increases). For the moderate participation group, instructional efficacy, leadership, reflective practice, and collaboration scores seem to be increasing while leadership stays roughly the same. The high participation group sees increases in leadership and reflective practice while most other constructs stay the same. It is difficult to discern group patterns by looking at the mean scores between groups over time. In some cases, scores dip in year 2, then reverse direction again in year 3. In other cases, scores seem to incrementally decrease or increase over time. Different groups at different times have higher or lower scores across constructs.

Table 21.

Mean Teacher Survey Construct Scores between Low, Moderate, and High Participation Groups over Time

		T1		T2		T3	
		Mean	(SD)	Mean	(SD)	Mean	(SD)
No/Low Participation[a]	Instructional Efficacy	4.02	(0.52)	3.99	(1.06)	3.87	(0.78)
	Collegiality	3.75	(0.64)	3.51	(1.04)	3.45	(0.83)
	Leadership	3.26	(1.08)	3.26	(1.15)	3.28	(1.21)
	Ongoing Learning	2.52	(0.92)	2.02	(0.90)	1.85	(1.00)
	Reflective Practice	-	-	1.33	(0.79)	1.49	(1.10)
	Collaboration	3.60	(1.60)	4.25	(1.51)	4.12	(1.20)
Moderate Participation[b]	Instructional Efficacy	4.00	(1.01)	4.27	(0.68)	4.41	(0.60)
	Collegiality	3.64	(0.63)	3.65	(0.91)	3.75	(0.82)
	Leadership	4.25	(0.91)	4.34	(1.09)	3.95	(1.18)
	Ongoing Learning	2.68	(1.19)	1.94	(0.78)	1.62	(0.47)
	Reflective Practice	-	-	1.55	(0.96)	2.67	(1.20)
	Collaboration	4.23	(1.28)	4.48	(1.31)	4.66	(1.26)
High Participation[c]	Instructional Efficacy	4.15	(0.66)	4.03	(0.51)	4.33	(0.45)
	Collegiality	3.91	(0.60)	3.97	(0.50)	4.11	(0.54)
	Leadership	3.99	(0.92)	4.19	(0.77)	4.62	(1.22)
	Ongoing Learning	3.11	(1.41)	1.75	(0.37)	1.94	(1.05)
	Reflective Practice	-	-	1.70	(0.74)	3.08	(1.25)
	Collaboration	4.23	(0.84)	4.74	(0.78)	4.38	(0.51)

Note. Instructional Efficacy and Collegiality are on a 5-point scale (1-5): Strongly Disagree, Moderately Disagree, Neutral, Moderately Agree, and Strongly Agree. Ongoing Learning is on a 7-point scale (1-7): 0 times, 1 time, 2 times, 3 times, 4 times, 5 times, and 6 times or more. Leadership, Reflective Practice, and Collaboration are on an 8-point scale (1-8): Never, Yearly, Semesterly, Quarterly/Bi-Monthly, Monthly, Bi-Weekly, Weekly, and Daily.

[a] Time 1: n=12; Time 2: n=24; Time 3: n=26

[b] Time 1: n=24; Time 2: n=27; Time 3: n=21

[c] Time 1: n=11; Time 2: n=12; Time 3: n=12

A multivariate analysis of variance (one-way MANOVA) is conducted at T1, T2, and T3 to better understand group differences in mean scores on survey constructs (instructional efficacy, collegiality, leadership, ongoing learning, reflective practice, and collaboration) (reflective practice data is only available at T2 and T3). Results from the a one-way MANOVA at T1 show that there is not a statistically significant difference between the low, moderate, and high participation groups on the combined dependent variable $F(10, 82) = 1.297, p = .246$; Pillai's $\Lambda = .273$; partial $\eta^2 = .137$. At T2 there is again, no statistically significant difference between the low, moderate, and high participation groups on the combined dependent variable $F(12, 98) = 1.710, p = .076$; Pillai's $\Lambda = .346$; partial $\eta^2 = .173$.

At T3 there is in fact a statistically significant difference between the low, moderate, and high participation groups on the combined dependent variable $F(12, 100) = 2.689, p = .004$; Pillai's $\Lambda = .488$; partial $\eta^2 = .244$. MANOVA results indicate that during the first two years of the ITQ program there are no statistically significant participation group differences on teacher effectiveness constructs (combined variable). However, by year 3 of the ITQ program, the differences between participation groups, on the combined dependent variables, is statistically significant.

Table 22 shows the results from the follow up univariate tests, with a Bonferroni correction (to reduce Type I Error that may be introduced because of unequal variance between groups), for T3. There is a statistically significant difference in participation group scores on instructional efficacy $F(2, 54) = 4.253, p = .019$; partial $\eta^2 = .136$.; leadership $F(2, 54) = 5.181, p = .009$; partial $\eta^2 = .161$.; and reflective practice $F(2, 54) = 9.993, p < .001$; partial $\eta^2 = .270$. Significant results for these three constructs indicate there are significant group differences on constructs between at least one combination of the three participation groups.

Table 22.
Results from Follow-up Univariate Analysis of Variance (ANOVA) Test on Teacher Survey Constructs for Low, Moderate, and High Participation Groups at T3, with a Bonferroni Correction

	Source	Sum of Squares (SS)	*df*	Mean Square	*F*
Spring 3 (T3)	Instructional Efficacy	3.82	2	1.91	4.25*
	Collegiality	3.60	2	1.80	2.94
	Leadership	15.29	2	7.65	5.18**
	Ongoing Learning	0.90	2	0.45	0.60
	Collaboration	2.91	2	1.46	1.12
	Reflective Practice	27.96	2	13.98	9.99***

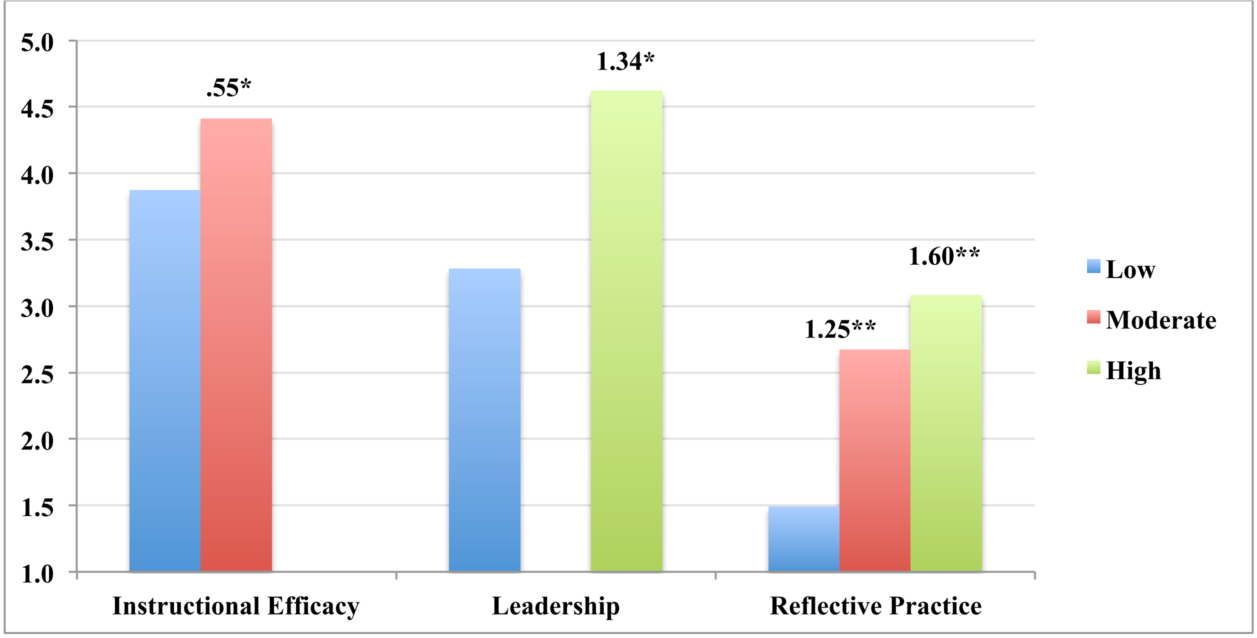
Games-Howell post-hoc tests (used to address unequal sample sizes and unequal variances) illuminate where group differences exist on instructional efficacy, leadership, and reflective practice (Table 23). For all three constructs, the moderate and high participation groups outscore the low participation group, suggesting participation in the ITQ program is related to higher scores on these aspects of teacher effectiveness. Games-Howell post-hoc tests show a statistically significant difference in instructional efficacy scores from 3.87 ± .78 in the low participation group to 4.41 ± .60 in the moderate participation group. The .55 mean difference is statistically significant (95% CI, .04 to 1.05) (*p* = .03). Games-Howell post-hoc tests show a statistically significant difference in leadership scores from 3.28 ± 1.21 in the low participation group to 4.62 ± 1.22 in the high participation group. The 1.34 mean difference is statistically significant (95% CI, .27 to 2.41) (*p* = .01). Additionally, Games-Howell post-hoc tests show a statistically significant difference in reflective practice scores from 1.49 ± 1.10 in the low participation group to 2.67 ± 1.20 in the moderate participation group, and from 1.49 ± 1.10 in the low participation group to 3.08 ± 1.25 in the high participation group. The 1.25 mean difference (moderate – low) (95% CI, .37 to 2.13) (*p* = .004), as well as the 1.60 mean difference (high – low) (95% CI, .53 to 2.66) (*p* = .003) are statistically significant.

Table 23.
Games-Howell Comparisons for Teacher Survey Constructs with Significant Group Differences between Low, Moderate, and High Participation Groups at T3

	(I)	(J)	(I-J)		95% Confidence Interval	
	Participation Group	Participation Group	Mean Diff	Std. Error	Lower Bound	Upper Bound
Instructional Efficacy	High	Moderate	-0.88	0.19	-0.56	0.39
	High	Low	0.46	0.20	-0.03	0.95
	Moderate	Low	0.55*	0.21	0.04	1.05
Leadership	High	Moderate	0.70	0.45	-0.43	1.82
	High	Low	1.34*	0.42	0.27	2.41
	Moderate	Low	0.64	0.37	-0.25	1.53
Reflective Practice	High	Moderate	0.35	0.46	-0.80	1.50
	High	Low	1.60**	0.42	0.53	2.66
	Moderate	Low	1.25**	0.36	0.37	2.13

Figure 4 illustrates the significant participation group differences for instructional efficacy, leadership, and reflective practice at T3. Findings from the MANOVA reveal the teacher survey's ability to detect group differences within a sample. The group differences discovered support the program theory by showing that teachers with greater levels of ITQP PD participation score higher on at least some of the teacher effectiveness constructs. The fact that there are no group differences early in the program and then significant group differences by the end of the program align with the ITQ Program theory. In the areas where there remains no group differences, it is unclear whether no group differences exist (failure of the program theory or failure of implementation), or if the measure itself is unable to detect differences. Regardless, the teacher survey demonstrates its ability to capture group differences on several constructs.

Figure 4.
Statistically Significant Group Differences on Teacher Survey Constructs between Low, Moderate, and High Participation Groups, T3



Testing for Between-Group Differences on the Teacher Survey

Given the significant group difference findings, it is also of interest to discover the extent to which participation group's experienced significant growth over time in their scores on teacher survey effectiveness constructs. A series of repeated measures ANOVAs, for each level of participation group, are used to detect significant change over time (T1, T2 , and T3) on teacher survey construct scores. 5 repeated measures ANOVAs are conducted to detect significant group differences over time for the low, moderate, and high participation groups. Additionally, a paired samples t-test is conducted between T2 and T3 for reflective practice, given the measure was only collected in year's 2 and 3.

Findings from the repeated measures ANOVAs indicate that there are no significant changes in the following construct scores for the low participation group over time: instructional efficacy, *F*(2, 58) = 1.248, *p* = .295; collegiality, *F*(2, 58) = .367, *p* = .694; leadership, *F*(2, 58) = .210, *p* = .811; or collaboration, *F*(2, 58) = .605, *p* = .550. Low participation group teachers do show statistically significant change over time in ongoing learning scores, *F*(2, 58) = 16.468, *p* = .0001, η^2 = .362, with ongoing learning scores steadily decreasing[7] from T1 (M = 3.01, SD = 1.24) to T2 (M = 2.06, SD = .81) to T3 (M = 1.77, SD = .89). Table 24 shows the results from the repeated measures ANOVAs for the low participation group. Post hoc analysis with a Bonferroni adjustment (to reduce Type I Error) reveals that ongoing learning scores statistically significantly decrease from T1 to T2 by .95 points (95% CI -1.57 to -.34) (*p* = .001), and from T1 to T3 by 1.24 points (95% CI -1.81 to -.68) (*p* = .0001). Pairwise comparisons are presented in Table 25.

Table 24

Results from Repeated Measures Analysis of Variance (ANOVA) Test on Teacher Survey Constructs over Time for No/Low Participation Teachers

Source	Sum of Squares (SS)	*df*	Mean Square	*F*
Instructional Efficacy	0.78	2	0.39	1.25
Collegiality	0.29	2	0.14	0.37
Leadership	0.21	2	0.10	0.21
Ongoing Learning	25.29	2	12.65	16.47***
Collaboration	0.75	2	0.37	0.61

Table 25

Pairwise Comparisons for Ongoing Learning with Significant Group Differences over Time, for No/Low Participation Teachers, with a Bonferroni Correction

(I)	(J)	(I-J)		95% Confidence Interval	
Time	Time	Mean Diff	Std. Error	Lower Bound	Upper Bound
T3	T2	-0.29	0.21	-0.84	0.25
T3	T1	-1.24***	0.22	-1.81	-0.68
T2	T1	-0.95**	0.24	-1.57	-0.34

[7] The change in Ongoing Learning scores is believed to be attributed to external contextual factors impacting the validity of the measure, which is discussed more fully in the Discussion section of this study. This is the case for all three participation groups.

A paired samples t-test (Table 26) for the no/low participation group on reflective practice reveals no statistically significant increase between T2 and T3. It appears that the ITQP theory of action is supported by findings from paired samples t-tests for the low group. For teachers minimally engaged in APD, no significant change is seen in reflective practice scores.

Table 26
Results from Paired Samples T-test on Reflective Practice between T2 and T3 for No/Low Participation Teachers

	T2	T3	*t*	*df*
Reflective Practice[a]	1.32 (0.80)	1.36 (0.82)	0.22	22

[a]Reflective Practice is on an 8-point scale (1-8): Never, Yearly, Semesterly, Quarterly/Bi-Monthly, Monthly, Bi-weekly, Weekly, and Daily.

Findings from the repeated measures ANOVAs indicate that there are no significant changes in the following construct scores for the moderate participation group over time: instructional efficacy, $F(2, 26) = .566$, $p = .574$; collegiality, $F(2, 26) = .365$, $p = .698$; leadership, $F(2, 26) = 1.798$, $p = .186$; or collaboration, $F(2, 22) = .308$, $p = .738$. Moderate participation group teachers do show statistically significant change over time in ongoing learning scores, $F(2, 24) = 10.369$, $p = .001$, $\eta^2 = .464$, with ongoing learning scores steadily decreasing from T1 (M = 3.04, SD = 1.19) to T2 (M = 2.19, SD = .92) to T3 (M = 1.75, SD = .48). Table 27 shows the results from the repeated measures ANOVAs for the moderate participation group. Post hoc analysis with a Bonferroni adjustment (to reduce Type I Error) reveals that ongoing learning scores statistically significantly decrease from T1 to T3 by 1.29 points (95% CI -2.15 to -.43) ($p = .004$). Pairwise comparisons are presented in Table 28.

Table 27.

Results from Repeated Measures Analysis of Variance (ANOVA) Test on Teacher Survey Constructs over Time for Moderate Participation Teachers

Source	Sum of Squares (SS)	*df*	Mean Square	*F*
Instructional Efficacy	0.32	2	0.20	0.57
Collegiality	0.41	2	0.20	0.37
Leadership	2.68	2	1.34	1.80
Ongoing Learning	11.14	2	5.57	10.37**
Collaboration	0.31	2	0.15	0.31

Table 28.

Pairwise Comparisons for Ongoing Learning with Significant Group Differences over Time, for Moderate Participation Teachers, with a Bonferroni Correction

(I)	(J)	(I-J)		95% Confidence Interval	
Time	Time	Mean Diff	Std. Error	Lower Bound	Upper Bound
T3	T2	-0.44	0.23	-1.09	0.21
T3	T1	-1.29**	0.31	-2.15	-0.43
T2	T1	-0.85	0.31	-1.71	0.02

A paired samples t-test (Table 29) for the moderate participation group on reflective practice reveals a statistically significant increase between T2 and T3, *t*(14) = 2.918, *p* = .011, *d* = .75. Reflective practice scores significantly increased 1.00 points between T2 and T3 (95% CI .26 to 1.74). Cohen's d indicates that the effect size is moderate. The increase in reflective practice scores for the moderate group is meaningful considering the ITQ program focus on Take One! in year 3 (which incorporates reflection on videotaped lessons).

Table 29.

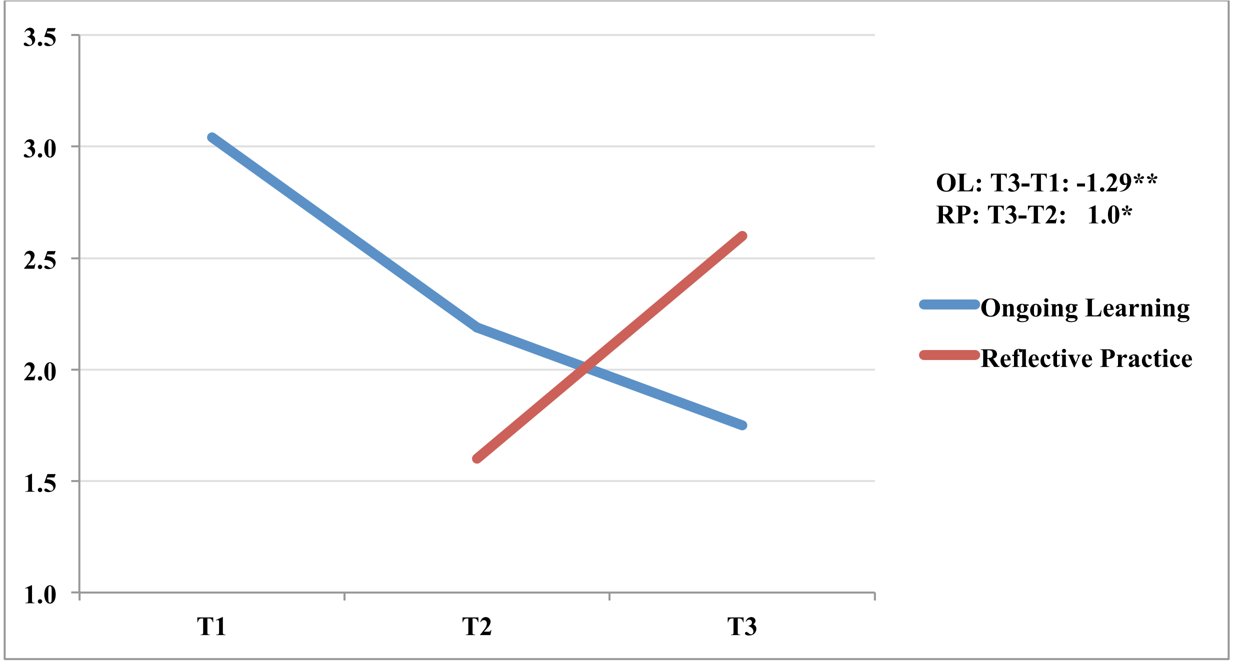
Results from Paired Samples T-test on Reflective Practice between T2 and T3 for Moderate Participation Teachers

	T2	T3	*t*	*df*
Reflective Practice	1.60	2.60	2.98*	14
	(1.06)	(1.37)		

[a]Reflective Practice is on an 8-point scale (1-8): Never, Yearly, Semesterly, Quarterly/Bi-Monthly, Monthly, Bi-weekly, Weekly, and Daily.

The lack of other substantive increases in teacher effectiveness for the moderate participation group is disappointing for the ITQ program and brings into question the participation group differences found in year 3, presented earlier. There is no significant growth in instructional efficacy for the moderate group, yet group differences exist at T3 between the low and moderate participation groups (these differences were nonexistent at T1). Repeated measures ANOVA findings do not fully substantiate earlier MANOVA findings leaving interpretation unclear. Figure 5 displays the two constructs for which the moderate participation group experienced significant growth over time.

Figure 5.
Significant Change over Time for the Moderate Participation Group on Ongoing Learning and Reflective Practice



A repeated measures analysis of variance discerns significant changes between T1, T2, and T3 on teacher survey effectiveness constructs for the high participation group. 5 repeated measures ANOVA's are conducted to detect significant group differences over time. A paired samples t-test is conducted between T2 and T3 for reflective practice, given the measure was only collected in year's 2 and 3.

Findings from the repeated measures ANOVAs indicate that there are no significant changes in the following construct scores for the high participation group over time: instructional efficacy, $F(2, 20) = 1.160$, $p = .334$; collegiality, $F(2, 20) = .931$, $p = .410$; leadership, $F(2, 20) = 2.881$, $p = .080$; or collaboration, $F(2, 20) = 2.170$, $p = .140$. High participation group teachers do show statistically significant change over time in ongoing learning scores, $F(2, 20) = 7.299$, $p = .004$, $\eta^2 = .422$, with ongoing learning scores decreasing from T1 (M = 3.31, SD = 1.41) to T2 (M = 1.75, SD = .37), and increasing from T2 to T3 (M = 1.93, SD = 1.10). Table 30 shows the results from the repeated measures ANOVAs.

Table 30.
Results from Repeated Measures Analysis of Variance (ANOVA) Test on Teacher Survey Constructs over Time for High Participation Teachers

Source	Sum of Squares (SS)	*df*	Mean Square	*F*
Instructional Efficacy	0.32	2	0.16	1.16
Collegiality	0.45	2	0.22	0.93
Leadership	1.21	2	0.61	2.88
Ongoing Learning	12.06	2	6.03	7.30**
Collaboration	1.45	2	0.72	2.17

Post hoc analysis with a Bonferroni adjustment (to reduce Type I Error) reveals that ongoing learning scores statistically significantly decrease from T1 to T2 by 1.36 points (95% CI -2.57 to -.15) ($p = .027$), and statistically significantly decrease from T1 to T3 by 1.18 points (95% CI -2.37 to -.04) ($p = .04$). Pairwise comparisons are presented in Table 31. The biggest drop in ongoing learning scores for the high participation group happened between year's 1 and 2, with a slight and insignificant increase in year 3.

Table 31.
Pairwise Comparisons for Ongoing Learning with Significant Group Differences over Time, for High Participation Teachers, with a Bonferroni Correction

(I)	(J)	(I-J)		95% Confidence Interval	
Time	Time	Mean Diff	Std. Error	Lower Bound	Upper Bound
T3	T2	0.18	0.34	-0.79	1.15
T3	T1	-1.18*	0.40	-2.33	-0.37
T2	T1	-1.36*	0.27	-2.57	-0.15

A paired samples t-test (Table 32) for the high participation group on reflective practice reveals a statistically significant increase between T2 and T3, *t*(9) = 2.56, *p* = .03, *d* = .81. Reflective practice scores significantly increased 1.00 points between T2 and T3 (95% CI .11 to 1.89). Cohen's d indicates that the effect size is moderately large. Considering the ITQ program focus on Take One! in year 3 (which incorporates reflection on videotaped lessons), it is not a surprise that there is a significant increase for the reflective practice construct.

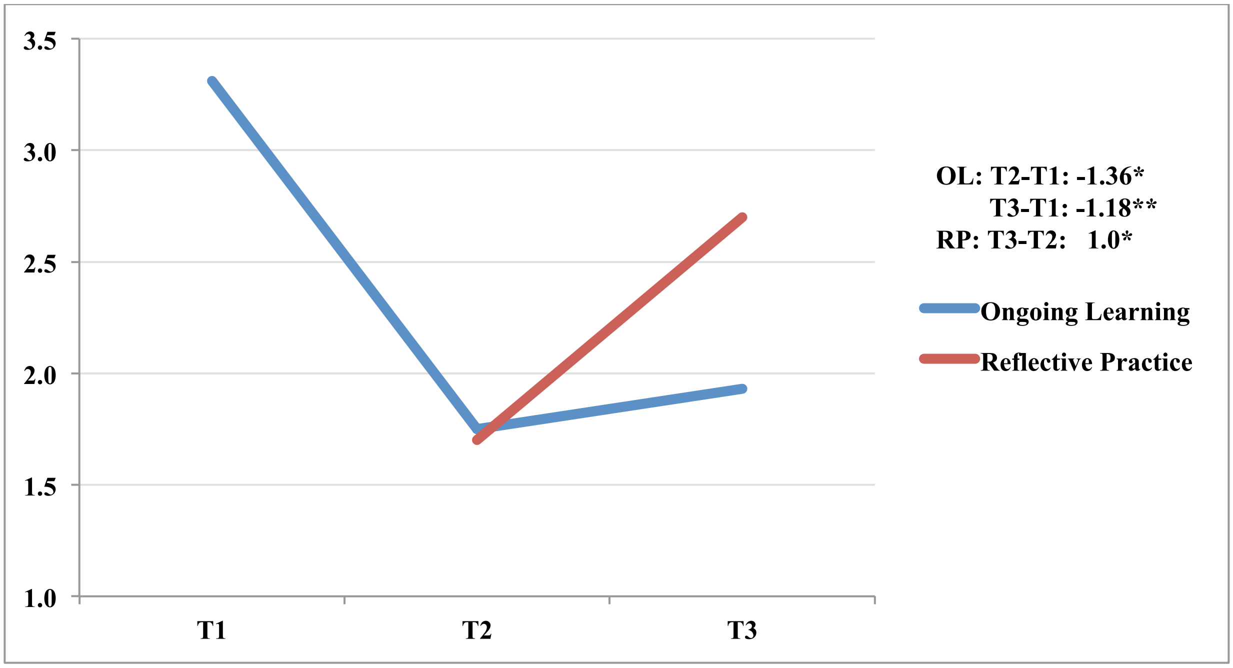
Table 32.
Results from Paired Samples T-test on Reflective Practice between T2 and T3 for High Participation Teachers

	T2	T3	*t*	*df*
Reflective Practice	1.70 (0.74)	2.70 (0.96)	2.56*	9

[a]Reflective Practice is on an 8-point scale (1-8): Never, Yearly, Semesterly, Quarterly/Bi-Monthly, Monthly, Bi-weekly, Weekly, and Daily.

The lack of much other substantive increases in teacher effectiveness for the high participation group is disappointing for the ITQ Program and confounds the interpretation of participation group differences found in year 3, presented earlier. There is no significant growth for the high participation group in leadership, yet group differences exist at T3 between the low and high groups on this construct. Without significant growth, the repeated measures ANOVA does not substantiate the significant participation group difference findings (MANOVA), leaving interpretation unclear. Figure 6 displays the two constructs for which the high participation group experienced significant change over time.

Figure 6.
Significant Change over Time for the High Participation Group Teachers on Ongoing Learning and Reflective Practice



Analysis of the teacher survey data sheds light on the measure's sensitivity to detecting differences, both between groups and over time. Findings reveal that all three groups experienced statistically significant decline in ongoing learning scores, at some point in the program. Furthermore, the moderate and high participation groups experienced significant growth in reflective practice scores. The teacher survey does exhibit a reasonable amount of sensitivity to capturing within-group differences for the ongoing learning and reflective practice constructs. However, the teacher survey's ability to capture between group differences is limited. At T3, the measure does discern significant participation group differences on 3/6 construct scores. If this were paired with matching within-group significant change over time (like it is for reflective practice scores), between-group differences would be more believable. Overall, for this sample, the teacher survey does not do a very good job of capturing within-group or between-group differences, limiting its use in a comprehensive teacher evaluation system.

Expert Assessment

There are three time points (Y2T2, Y3T1, Y3T2) available for analysis with the expert assessment data. A mixed ANOVA is run to determine whether there are differences between participation groups (low, moderate, high) over time (T1, T2, T3) on their overall strategy use (DV). Overall strategy use is an average of ratings on reading, writing, inquiry, collaborative, and other strategies. Given there are some expected fluctuations in type of strategy use between content areas and the program valuing of various instructional strategies, looking at overall strategy use is appropriate. The mixed ANOVA statistical test helps determine if overall strategy use at the different time points is dependent on PD participation group. The expectation is that the high participation group will have higher ratings and that these ratings will increase at a greater rate (than other participation groups) over time. For this analysis, there is data available at three time points for 20 teachers in the no/low participation group, 16 teachers in the moderate participation group, and 12 teachers in the high participation group. Mixed ANOVA results indicate that there is a statistically significant interaction between group and time on overall strategy use, $F(4, 90) = 4.575$, $p = .003$, partial $\eta^2 = .169$.

An exploration of the simple main effects for time illuminates the difference between groups at each time level. Table 33 shows the results from the standard one-way ANOVAs on overall strategy use at T1, T2, and T3. There is a statistically significant difference in overall strategy use between groups at all three time points: T1, $F(2, 45) = 6.131$, $p< .001$, partial $\eta^2 = .368$; T2, $F(2, 48) = 7.492$, $p< .001$, partial $\eta^2 = .410$; and T3, $F(2, 48) = 9.128$, $p< .001$, partial $\eta^2 = .388$. Being that T1 of the expert assessment corresponds to the spring of program year 2, some level of program effect is expected. There are in fact statistically significant differences in mean scores for overall strategy use between at least two of the three groups at each time point.

Table 33.

Results from Standard One-Way ANOVA Test of Between-Subjects Effect (Participation Group) on Overall Strategy Use Rating at Y2T2 (T1), Y3T1 (T2), and Y3T2 (T3)

	Source	Sum of Squares (SS)	*df*	Mean Square	*F*
	Between	12.26	2	6.13	13.10^{***}
T1	Error	21.07	45	0.47	
	Total	202.48	48		
	Between	14.98	2	7.49	16.68^{***}
T2	Error	21.56	48	0.45	
	Total	230.28	51		
	Between	18.26	2	9.13	15.23^{***}
T3	Error	28.77	48	0.60	
	Total	344.64	51		

Results from the univariate analysis of variance are presented in Table 34. Tukey post-hoc tests show a statistically significant difference in overall strategy use between the high and moderate and high and low groups at T1 and T2. At T3, there is a significant difference in overall strategy use between all three groups. At T1, the mean score for overall strategy use for the high participation group is 1.15 points higher than the low group, and 1.19 points higher than the moderate group ($p < .001$ for both). Similarly, at T2, the mean score for overall strategy use for the high participation group is 1.35 points higher than that for the low group, and 1.17 points higher than that for the moderate group ($p < .001$ for both). At T3 the moderate group mean score on overall strategy use is .61 points higher than the mean score for the low group (*p* < .05). At T3, the high participation group mean score on overall strategy use is 1.55 points higher than that for the low group (*p* < .001). While the mean difference increases over time between the high and low participation groups and between the moderate and low participation groups, the mean difference decreases over time between the high and moderate groups. At T3, the mean overall strategy use score for the high participation group is .94 points higher than the mean score for the moderate participation group (*p* < .01). Given the four-point scale on the expert assessment, mean differences are not only statistically significant but also practically significant.

Table 34.

Results from Tukey HSD Comparisons for Overall Strategy Use Rating between No/Low, Moderate, and High Participation Groups at Y2T2 (T1), Y3T1 (T2), and Y3T2 (T3)

	(I)	(J)	(I-J)		95% Confidence Interval	
	Participation Group	Participation Group	Mean Diff	Std. Error	Lower Bound	Upper Bound
T1	High	No/Low	1.15***	0.25	0.54	1.76
	High	Moderate	1.19***	0.26	0.55	1.82
	Moderate	No/Low	-0.04	0.23	-0.59	0.52
T2	High	No/Low	1.35***	0.24	0.76	1.93
	High	Moderate	1.17***	0.25	0.57	1.78
	Moderate	No/Low	0.17	0.22	-0.35	0.69
T3	High	No/Low	1.55***	0.28	0.87	2.22
	High	Moderate	0.94**	0.29	0.24	1.64
	Moderate	No/Low	0.61*	0.25	0.00	1.21

Mean overall strategy rating for all three groups over time is presented in Table 35. At T1, overall strategy use increased from the moderate participation group (*M* = 1.56, *SD* = .56), to the no/low participation group (*M* = 1.60, *SD* = .59), to the high participation group (*M* = 2.75,*SD* = .95), in that order. At T2, overall strategy use increased from the low participation group (*M* = 1.53, *SD* = .47), to the moderate participation group (*M* = 1.63, *SD* = .67), to the high participation group (*M* = 2.92, *SD* = .75). Finally, at T3, the differences are most extreme, with overall strategy use increasing from the low participation group (*M* = 1.81, *SD* = .77), to the moderate participation group (*M* = 2.38, *SD* = .79), to the high participation group (*M* = 3.38, *SD* = .73). Both higher participation groups are rated as using various instructional strategies at statistically significantly higher rates than the low participation group by the end of the ITQ program.

Table 35.

Mean Scores for Overall Strategy Use between Participation Groups over Time

		T1		T2		T3	
	n	M	(SD)	M	(SD)	M	(SD)
No/Low Participation Group	20	1.60	(0.59)	1.53	(0.47)	1.81	(0.77)
Moderate Participation Group	16	1.56	(0.56)	1.63	(0.67)	2.38	(0.79)
High Participation Group	12	2.75	(0.95)	2.92	(0.75)	3.38	(0.73)

Note. The Expert Assessment is on a 4-point scale (1-4): Barely/Never, Sometimes, Often, and Most of the time/Always.

Exploration of the simple main effects for time reveals the extent to which each level of participation group experiences statistically significant growth in overall strategy use over time. Table 36 displays the Huynh-Feldt (used to address violation of sphericity when epsilon > .75 (Field, 2012)) results from the three repeated measures ANOVAs. For the no/low participation group, time does not have a statistically significant effect on overall strategy use, *F*(1.39, 26.39) = 3.88, *p* = .05, partial η^2 = .17. For both the moderate and high participation groups, there is a statistically significant effect of time on overall strategy use (moderate: *F*(1.71, 25.68) = 3.88, *p* < .001, partial η^2 = .65; high: *F*(1.29, 22) = 14.82, *p* < .001, partial η^2 = .57). Estimates of effect size for both the moderate and high participation groups are relatively high indicating the effect is likely practically significant as well.

Table 36.

Results from Huynh-Feldt ANOVA Tests of Within-Subjects Effect (Time) on Overall Strategy Use for Low, Moderate, and High Participation Groups

Group	Sum of Squares (SS)	*df*	Mean Square	*F*
No/Low Participation	0.85	1.34	0.61	3.87
Error	4.16	26.39	0.16	
Moderate Participation	6.54	1.71	3.82	27.94***
Error	3.51	25.68	0.14	
High Participation	2.59	2	1.29	14.82***
Error	1.92	22	0.09	

Close inspection of the pairwise comparison table reveals where the statistically significant differences lie for the moderate and high participation groups (Table 37). Both the moderate and high participation groups experience significant growth over time in their scores on overall strategy use between T1 and T3. This translates to significant growth between Spring of program year 2 and Spring of program year 3. During this time, the moderate participation group gains .81 points ($p < .001$) on their overall strategy use score and the high participation group gains .63 ($p < .01$) points on their overall strategy use score. Both the moderate and high participation groups also experience significant growth in overall strategy use score between T2 and T3. Between the Fall of program year 3 and Spring of program year 3, the moderate participation group gains .75 points ($p < .001$) on their overall strategy use score and the high participation group gains .47 points ($p < .01$) on their overall strategy use score. The greatest gains are made from one year to the next and the most drastic gains are experienced within the moderate participation group. Given the 4-point scale, gains for each group are not only statistically significant but also practically significant. There is no statistically significant growth between T1 and T2 for either group.

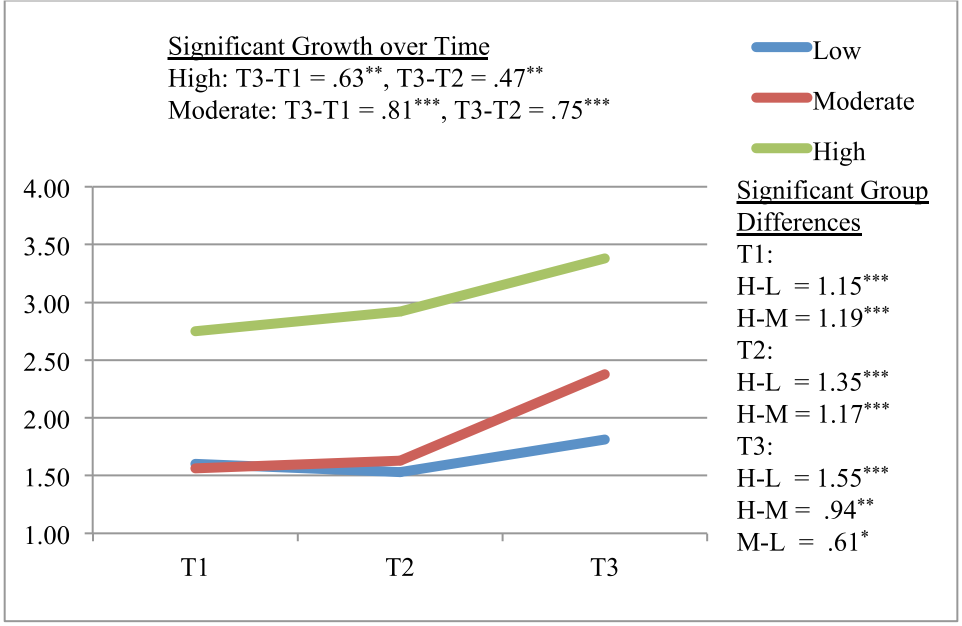
Table 37.

Results from Pairwise Comparisons for Within-Subjects Effect (Time) on Overall Strategy Use for No/Low, Moderate, and High Participation Groups, with a Bonferroni Adjustment

	(I)	(J)	(I-J)		95% Confidence Interval	
	Participation Group	Participation Group	Mean Diff	Std. Error	Lower Bound	Upper Bound
	T3	T1	0.21	0.12	-0.10	0.52
No/Low Participation	T3	T2	0.28	0.13	-0.05	0.61
	T2	T1	-0.07	0.06	-0.22	0.08
	T3	T1	0.81^{***}	0.12	0.50	1.13
Moderate Participation	T3	T2	0.75^{***}	0.15	0.32	1.15
	T2	T1	0.06	0.09	-0.19	0.31
	T3	T1	0.63^{**}	0.14	0.24	1.03
High Participation	T3	T2	0.47^{**}	0.10	0.18	0.76
	T2	T1	0.17	0.12	-0.17	0.50

Figure 7 displays mean scores on overall strategy use over time for each participation group and summarizes both the statistically significant mean group differences and the statistically significant growth for each group. Findings from the mixed ANOVA make known the expert assessment's ability to detect both changes over time (within-group) and participation group differences on construct scores (between-groups). The expert assessment used in this study is in fact sensitive to capturing these differences both across years and within a year's time. Given this instrument's apparent high level of sensitivity in these regards, it is unfortunate that the available data does not extend to year 1 of ITQP implementation (or, prior). It is highly likely that even greater differences could have been captured.

Figure 7.
Statistically Significant Growth and Mean Group Difference on Overall Strategy Use



Classroom Observations

Out of the 32 teachers for which classroom observation data is available, only half (16) agreed to participate in two classroom observations (one per semester in year three). 5 of these teachers belong to the moderate participation group and 11 belong to the high participation group. For these 16 teachers, I am able to conduct analyses looking at change over time. Given that the program expects changes in ratings over time for these folks, especially for the high-participation group, I look at change over time separately for these two groups. To test for change over time, a series of paired sample t-tests are run on classroom environment (including 4 sub-constructs), instruction (including 5 sub-constructs), and total strategy usage (including 5 sub-constructs). Tables 9 and 10, presented earlier, list instructional strategies as well as the standards and components used in this set of analyses.

Results from the paired samples t-tests reveal that for the moderate participation group, there are no significant changes in scores on classroom environment (and sub-constructs), instruction (and sub-constructs), or total instructional strategy usage (and sub-constructs) between the Fall and Spring observations. Table 38 shows the results for the moderate participation group.

Table 38.
Results from Paired Samples T-tests for Moderate Participation Teachers on Classroom Observation Constructs and Sub-Constructs between Y3T1 (Fall 2011) and Y3T2 (Spring 2012)

	Y3T1		Y3T2			
	M	(SD)	M	(SD)	*t*	*df*
Classroom Environment	2.74	(0.47)	2.33	(0.90)	-1.71	4
S2CA	3.07	(0.49)	2.40	(1.09)	-2.00	4
S2CB	2.70	(0.60)	2.25	(0.53)	-2.71	4
S2CC	2.33	(0.62)	2.13	(1.12)	-0.41	4
S2CD	2.87	(0.69)	2.53	(0.96)	-1.58	4
Instruction	2.16	(0.46)	2.02	(0.61)	-0.70	4
S3CA	2.25	(0.64)	2.05	(0.86)	-0.83	4
S3CB	1.87	(0.61)	1.67	(0.53)	-0.56	4
S3CC	2.15	(0.45)	2.25	(0.64)	0.43	4
S3CD	2.05	(0.48)	1.95	(0.60)	-0.78	4
S3CE	2.50	(0.50)	2.20	(0.57)	-1.18	4
Total Strategy Usage	6.00	(1.00)	4.80	(2.68)	-0.89	4
Reading	1.00	(0.71)	2.00	(2.12)	0.88	4
Writing	1.80	(0.84)	1.20	(0.84)	-1.18	4
Inquiry	1.20	(0.45)	1.20	(0.84)	0.00	4
Collaborative	1.20	(1.10)	0.40	(0.55)	-1.37	4
Other	0.80	(0.84)	0.00	(0.00)	-2.14	4

Note. The Classroom Observation Protocol is on a 4-point scale (1-4): Ineffective, Developing, Effective, and Highly Effective. Mean scores on Instructional Strategies reflect an actual count of distinct Instructional Strategies used per class period.

Results from a series of paired samples t-tests reveals that for the high participation group, there are statistically significant gains in both the classroom environment and instruction construct and/or sub-construct scores. Table 39 shows the results from the paired samples t-tests for this group. The high-participation teachers see an increase in mean scores between Y3T1 and Y3T2 scores: (S2) classroom environment scores increased by .23 points (95% CI, .1163 to .3420) from Y3T1 to Y3T2 *t*(10) = 4.525, *p* = .001, *d* = 1.36; (S2CA) creating an environment of respect and rapport scores increased by .27 points (95% CI, .0770 to .4684) from Y3T1 to Y3T2 *t*(10) = 3.105, *p* = .011, *d* = .94; (S2CD) managing student behavior scores increased by .48 points (95% CI, .2322 to .7375) from Y3T1 to Y3T2 *t*(10) = 4.276, *p* = .002, *d* = 1.29; and (S3CD) using assessment in instruction to advance student learning scores increased .39 points (95% CI, .1565 to .6162) from Y3T1 to Y3T2 *t*(10) = 3.746, *p* = .004, *d* = 1.13. Practically speaking, all 4 increases in classroom environment and instruction are notable, given the 4-point scale and the relatively short time elapsed between observations. There are no statistically significant differences in total strategy usage (including sub-constructs) between the Fall and Spring observations.

Findings from the paired samples t-tests conducted with the classroom observation data point to the fact that for teachers in which one expects to see significant changes due to intensive participation in targeted professional development, the classroom observation protocol is in fact sensitive to capturing changes. For teachers that engage in more moderate levels of PD, the classroom observation protocol is not sensitive enough to capture changes in a given academic year. This may be to the lack of significant changes in these folks or due to the bigger picture constructs that are measured with this instrument. Due to the single year administration, it is unclear as to whether classroom observations can detect changes in teacher effectiveness from one year to the next. It is likely that given a larger range of time, one would see sensitivity levels increase.

Table 39.

Results from Paired Samples T-tests for High Participation Teachers on Classroom Observation Constructs and Sub-Constructs between Y3T1 (Fall 2011) and Y3T2 (Spring 2012)

	Y3T1		Y3T2			
	M	(SD)	M	(SD)	*t*	*df*
Classroom Environment	2.41	(0.85)	2.64	(0.87)	4.53^{**}	10
S2CA	2.36	(0.92)	2.64	(0.96)	3.11^{*}	10
S2CB	2.55	(0.78)	2.61	(0.84)	0.43	10
S2CC	2.52	(0.94)	2.61	(0.93)	0.61	10
S2CD	2.21	(0.92)	2.70	(0.90)	4.28^{**}	10
Instruction	2.23	(0.92)	2.40	(0.79)	0.13	10
S3CA	2.39	(0.97)	2.52	(0.88)	0.84	10
S3CB	2.21	(0.91)	2.27	(0.92)	0.33	10
S3CC	2.27	(0.93)	2.43	(0.68)	1.02	10
S3CD	1.95	(0.81)	2.34	(0.70)	3.75^{**}	10
S3CE	2.32	(1.12)	2.41	(0.94)	0.56	10
Total Strategy Usage	6.55	(4.48)	5.82	(2.96)	-0.82	10
Reading	1.91	(1.64)	1.64	(1.21)	-1.00	10
Writing	1.45	(1.75)	1.64	(0.92)	0.41	10
Inquiry	1.73	(1.56)	1.45	(0.93)	-0.71	10
Collaborative	1.09	(0.94)	0.73	(0.79)	-1.49	10
Other	0.36	(0.51)	0.36	(0.51)	0.00	10

Note. The Classroom Observation Protocol is on a 4-point scale (1-4): Ineffective, Developing, Effective, and Highly Effective. Mean scores on Instructional Strategies reflect an actual count of distinct Instructional Strategies used per class period.

Given the relatively few constructs for which there is significant change over time, for cases in which there are 2 observations, average scores are used for subsequent analyses. This allows for more accurate comparisons to be made with teachers that only have 1 time-point available for analysis that may be in either the Fall or Spring semester. This decision is supported by the literature indicating that multiple estimates create a more valid portrayal of effectiveness (Danielson, 2007; LAUSD, 2013; MET Project, 2013; UTLA, 2013).

A series of one-way ANOVAs are conducted on classroom environment and instruction to test for significant group differences. One-way ANOVAs reveal no statistically significant differences in classroom environment, or instruction, or, corresponding sub-constructs. Table 40 displays mean scores on classroom environment and instruction and their underlying components. Testing for group differences between low, moderate, and high participation groups produces no significant findings. Average scores hover between the developing (2) to effective (3) range for all groups. Classroom environment and its components consistently receive higher scores than instruction and its components. The data shows that delivery of instruction is no better than the environment in which it occurs.

Table 40.
Mean Scores on Standards and Components between Low, Moderate, and High Participation Groups

	No/Low		Moderate		High	
	n=6		n=12		n=14	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
Classroom Environment	2.97	(.52)	2.47	(.56)	2.54	(.80)
S2CA	3.22	(.50)	2.64	(.64)	2.54	(.85)
S2CB	2.92	(.56)	2.39	(.54)	2.58	(.72)
S2CC	2.83	(.59)	2.43	(.72)	2.58	(.88)
S2CD	2.89	(.69)	2.43	(.67)	2.48	(.83)
Instruction	2.46	(.59)	2.11	(.44)	2.30	(.77)
S3CA	2.50	(.77)	2.25	(.56)	2.39	(.81)
S3CB	2.50	(.59)	1.93	(.43)	2.24	(.82)
S3CC	2.54	(.46)	2.19	(.52)	2.33	(.70)
S3CD	2.42	(.51)	2.02	(.68)	2.19	(.68)
S3CE	2.33	(.75)	2.15	(.45)	2.29	(.96)

Note. The Classroom Observation Protocol is on a 4-point scale (1-4): Ineffective, Developing, Effective, and Highly Effective.

A MANOVA[8] is used to determine if there are participation group differences on total strategy usage. The differences between participation groups on the combined dependent instructional strategy variables is not statistically significant, $F(10, 50) = .477$, $p = .897$; Wilks' $\Lambda = .833$; partial $\eta^2 = .087$. Table 41 displays mean scores on instructional strategy use between low, moderate, and high participation groups. On average, teachers use between 5.3 (moderate)

[8] A one-way MANOVA using the 3 constructs of classroom environment, instruction, and instructional strategy usage is not possible due to the incidence of multicollinearity between all three constructs. Similarly, one-way MANOVAs using the sub-constructs under classroom environment and instruction are not possible due to issues with multicollinearity.

to 6 (high) to 6.8 (low) different instructional strategies per class period. Reading strategies are the most commonly used across groups while other strategies are the least used across participation groups. Although there appear to be some differences between groups on mean strategy use, none of the differences are statistically significant.

Table 41.
Mean Usage of Instructional Strategies between Low, Moderate, and High Participation Groups

	No/Low		Moderate		High	
	n=6		n=12		n=14	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
Total Strategy Use	6.83	(3.97)	5.33	(2.71)	6.00	(3.13)
Reading	2.33	(1.75)	1.46	(1.01)	1.82	(1.27)
Writing	1.17	(0.75)	1.29	(0.66)	1.43	(1.14)
Inquiry	1.83	(1.33)	1.33	(0.78)	1.61	(1.15)
Collaborative	0.83	(0.98)	0.75	(0.66)	0.79	(0.75)
Other	0.67	(0.82)	0.50	(0.88)	0.36	(0.36)

Note. Mean scores reflect an actual count of distinct Instructional Strategies used per class period.

The classroom observation protocol does not seem to be sensitive to detecting participation group differences across constructs/sub-constructs in this sample. It may be that there is in fact little difference in observable classroom practice between teachers at the school sites. One difficulty here is that overall, regardless of participation level, teachers are rated relatively low (below effective) providing a restricted range to detect differences. This may be due to overall teacher quality at these lower performing schools and/or it could be due to contextual factors impacting teacher effectiveness. However, given the drastic differences in level of PD participation, it is disappointing that the classroom observation protocol does not detect group differences between participation groups. Unfortunately, there is no baseline data available for this instrument, which could be used to better understand both change over time and participation group differences.

Summary Across Measures

The measures of teacher effectiveness used in this study are able to capture within-group differences (change over time) to varying extents. The teacher survey captures changes (negative) in ongoing learning scores for all three levels of the participation groups over time. For the moderate and high participation groups, changes (positive) in reflective practice are also captured over time. For the other 4 teacher effectiveness constructs, the teacher survey is not able to detect significant differences over time. The expert assessment does a good job of capturing within-group differences (all positive) both between one Spring to the next and from Fall to Spring. The classroom observation protocol does a moderately good job of detecting change over time in classroom environment and instruction (constructs and sub-constructs) for the high participation group. Given the relatively short period of time between observations (semester 1 to semester 2), the sensitivity displayed for the high participation group (where one would expect to see change) is promising. Given additional years of data, it is highly probable that additional change over time would be captured. The expert assessment measure of teacher effectiveness proves itself to be the most sensitive to detecting within-group differences in this study.

The measures of teacher effectiveness used in this study are less effective at capturing between-group differences. The teacher survey only captures significant group differences at T3 (end of the ITQ program) on instructional efficacy, leadership, and reflective practice. While higher participation groups do outscore the no/low participation group, the fact that the detected differences do not fully align (save reflective practice) with significant growth over time for either participation group on the same constructs calls into question the validity of the findings. It seems there may be other factors influencing construct scores (threatening internal validity of the instrument). The expert assessment proves to be the most sensitive to detecting between-group differences. At all three time-points significant differences on overall strategy usage are found between low, moderate, and high participation groups. The classroom observation protocol fails at detecting any between-group differences in year 3 of the study. Overall, few meaningful differences in scores on teacher effectiveness constructs are discovered between low, moderate, and high participation group teachers. One possible explanation here is that there are minimal group differences within the sample and the findings reflect less on the measures themselves than the teachers included in the study. Across the board, the expert assessment proves itself to be the most sensitive to detecting both within-group and between-group differences.

Research Question 2

The following research question and sub-questions are addressed in this section: (2.) What are the relationships between measures of teacher effectiveness and teacher characteristics? (a.) What is the strength and direction of association between variables? (b.) In what ways are teacher characteristics predictive of teacher effectiveness? (c.) In what ways are academic PD participation (APD) and/or total strategy use predictive of teacher effectiveness?

In order to answer the second research question regarding relationships between variables included in the study, I begin by looking at the continuous variables of interest, namely the 10 main constructs located across instruments and additionally total academic PD participation in ITQP. Table 42 displays the Pearson correlation coefficients between the 11 continuous variables at Y3T2. Given that all 10 constructs are measures of teacher effectiveness, a positive association between variables is expected. Given that ITQP was designed to increase teacher effectiveness, it is expected that APD is also positively associated with all 10 constructs. Results from testing of assumptions are presented in Appendix G.

Table 42.
Pearson Correlation Coefficient Table of Continuous Variables of Interest used in this Study, Y3T2

Instrument	Teacher Survey						EA	Classroom Obs.			APD
	Instructional Efficacy	Collegiality	Leadership	Ongoing Learning	Reflective Practice	Collaboration	Overall Strategy Use	Total Strategy Use	Classroom Environment	Instruction	Total APD
Instructional Efficacy	1.00										
Collegiality	0.26*	1.00									
Leadership	0.41**	0.25	1.00								
Ongoing Learning	0.18	0.15	0.41**	1.00							
Reflective Practice	0.51***	0.21	0.44**	0.24	1.00						
Collaboration	0.31*	0.30*	0.27*	0.21	0.35**	1.00					
Overall Strategy Use	0.40**	0.36*	0.42**	0.18	0.41**	0.01	1.00				
Total Strategy Use	0.24	-0.19	0.14	0.23	0.16	-0.11	0.01	1.00			
Classroom Environment	0.30	-0.16	0.05	0.19	0.04	-0.12	0.11	0.78***	1.00		
Instruction	0.33	-0.10	0.16	0.21	0.24	-0.03	0.23	0.83***	0.94***	1.00	
Total APD	0.34**	0.34**	0.40**	0.03	0.50***	0.16	0.61***	-0.04	-0.21	-0.04	1.00

Looking at the teacher survey constructs, it is clear that all six constructs are positively correlated with each other, essentially supporting the claim that these are all measures of teacher effectiveness. Teachers showing high effectiveness in one area (measured by the teacher survey) generally score high on the other teacher effectiveness constructs. In several places, these measures are statistically significantly correlated with each other (most often at a moderate level). Instructional efficacy, leadership, and collaboration are statistically significantly correlated with 4/5 constructs on the teacher survey, overall strategy use from the expert assessment, and total academic PD participation. Reflective practice mirrors the same relationships with 3/5 constructs on the teacher survey and overall strategy use and APD. Collegiality is statistically significantly correlated with instructional efficacy, collaboration, overall strategy use, and APD. Ongoing learning shows a weak to negligible positive correlation to all constructs across instruments, except for a moderate positive statistically significant correlation with leadership. The expert assessment data is primarily positively and statistically significantly correlated with teacher survey constructs.

Instructional efficacy, collegiality, leadership, reflective practice, and overall strategy use are the most heavily associated with ITQP APD participation. Collaboration and ongoing learning are not statistically significantly correlated with APD suggesting that ITQP may not be associated with change in these areas. Academic PD participation, while typically positively correlated with teacher effectiveness constructs, demonstrates just the opposite relationship with the classroom observation data (although not significant). It remains a question as to why this negative relationship is present. Classroom observation teacher effectiveness constructs exhibit a very strong positive correlation with one another. Interestingly, the classroom observation data behaves differently with the other measures of teacher effectiveness and APD. It seems that unique information is generated by the classroom observation data. The lack of correlation with other measures both generates questions about triangulation and points to its importance for inclusion in a comprehensive teacher evaluation system.

Exploration into relationships between independent variables and dependent variables is pursued with regression analysis. Looking at the regression coefficient for each of the independent variables for each of the dependent variables gives an idea of how each variable behaves. All independent variables are dichotomous, coded 0/1, with 1 indicating the hypothesized positive association with teacher effectiveness. For example, it is expected that a teacher with a clear credential (0 = No, 1 = Yes) will have higher scores on teacher effectiveness measures because of the higher level of certification/qualification (input).

Table 43 provides a summary of the signs and strengths of the regression coefficients found across all the final regression models (forthcoming). Surprisingly, the regression coefficient for clear credential is always negative (and in one place significant). This could be due to the lower level of support teachers receive once they finish the new teacher induction program. School B always has a positive regression coefficient (significant in a few places) signifying the hypothesis is most likely correct (school B had less turnover and chaos than school A therefore was expected to perform better). Interestingly, in very few places do teacher characteristics behave the same way across all constructs. Regression coefficients for full-time teacher are often positive, indicating predicted higher scores on teacher effectiveness measures. Regression coefficients for established and content SS are very mixed indicating an uncertain association with teacher effectiveness measures. In this study, teacher characteristics do not provide enough information to accurately predict teacher effectiveness. For only two of the constructs; ongoing learning and collaboration - are the teacher characteristics as a whole significant in predicting scores on constructs (to be discussed). There remain other, more significant, factors contributing to scores on teacher effectiveness constructs.

Table 43.
Regression Coefficients and Statistical Significance according to Teacher Effectiveness Construct

Construct	Clear	Established	School B	Content SS	Full-Time Teacher
Instructional Efficacy	-0.38	-0.27	0.38*	0.18	0.11
Collegiality	-0.24	-0.03	0.24	-0.02	-0.32
Leadership	-0.38	0.07	0.29	-0.12	0.25
Ongoing Learning	-2.40***	0.18	0.17	-0.10	0.43
Reflective Practice	-0.43	-0.30	0.44	-0.01	0.48
Collaboration	-0.64	-0.17	1.03**	0.17	-0.02
Overall Strategy Use	-0.11	-0.21	0.20	-0.54*	0.32
Classroom Environment	-0.11	0.07	0.38	0.22	0.48
Instruction	-0.33	-0.03	0.54**	0.03	0.25

Exploring Predictive Power

Table 44 lists the frequency and percentage distribution for each of the independent variables considered for inclusion in the analyses for research question 2. In order to be kept in the analyses, a minimum amount of 10% variation needs to be present in the sample (this ensures necessary power and minimizes error introduced into the models. This excludes qualified to teach, national board certified, and veteran teacher from the analyses. Clear and BTSA are not needed in the same model because one voids the other; i.e. a clear credentialed teacher (input) will no longer be eligible for BTSA support (intervention). Therefore, the variables of clear credential, established at school site, school B, content SS, and full-time teaching status are used as predictor variables pertaining to teacher characteristics (reflecting inputs, experience, and school context).

Table 44.
Frequency and Percentage Distribution of Independent Variables at T3

	Yes		No	
	Frequency	Percent of the Total	Frequency	Percent of the Total
Clear	52	88	7	12
Qualified to Teach	55	93	4	7
National Board Certified	0	0	37	100
Veteran Teacher	56	95	3	5
Established Teacher	42	70	17	30
BTSA Support	7	12	52	88
School B	39	66	20	44
Content SS	30	51	29	49
Full-Time Teacher	53	90	6	10

A series of hierarchical multiple regression analyses are run in SPSS, allowing me to control for the effects of the covariates (teacher characteristics pertaining to inputs, experience, and school context) and determine the predictive power of academic PD participation (for teacher survey and expert assessment constructs) or total strategies (for classroom observation constructs) on the dependent variables of interest[9]. Where neither APD nor total strategy usage are significantly correlated with teacher effectiveness constructs (see Table 42, above) I test simply the predictive power of teacher characteristics on teacher effectiveness constructs (ongoing learning and collaboration) using multiple regression. Year 3 factor scores generated by a confirmatory factor analysis (where the sample mean is equal to zero) are used as outcome variables for the teacher survey, which provides a standardized way to make comparisons across constructs. For all regression models, assumptions are tested (independence of cases, linearity, and homoscedasticity) and adequately satisfied – results reported in Appendix G.

[9] For all hierarchical regression models, the percentage of the variance attributed to teacher characteristics and APD/or total strategy use are reported separately. It is recognized that the percentage of variance attributed to each block in the final model is contingent upon the order in which each block is entered into the model. Percentages are therefore not absolute but instead contingent on the structure of the final models.

I begin this set of analyses with the teacher effectiveness constructs housed within the teacher survey. Beginning with instructional efficacy, a hierarchical regression model analysis is conducted using 5 covariates in block one and APD in block two (Table 45). Model 2 is statistically significant, $R^2 = .285$, $F(6, 52) = 3.463$, $p = .006$; adjusted $R^2 = .203$. The addition of APD to the model leads to a statistically significant increase in R^2 of .183, $F(1, 52) = 13.303$, $p = .001$. The overall model accounts for 29% of the variation in in instructional efficacy, with 18% attributed strictly to APD (11% attributed to teacher characteristics). Model 2 produces the following equation: Instructional Efficacy = -.070 – (.379 x Clear) – (.274 x Established) + (.385 x School B) + (.180 x Content SS) + (.110 x Full-Time) + (.007 x APD); where a .5 SD increase in instructional efficacy is expected from 72 hours (beginning of the high participation group) of APD.

Table 45.
Hierarchical Multiple Regression Analysis Predicting Instructional Efficacy Score from Clear Credential, Established at School Site, School, Content Area, Full-Time Teaching Status, and Academic PD Total Hours of Participation

	Instructional Efficacy			
	Model 1		Model 2	
	B	β	B	β
Constant	0.118		-0.070	
Clear Credential	-0.419	-0.240	-0.379	-0.217
Established	-0.208	-0.167	-0.274	-0.220
School B	0.389^{*}	0.326	0.385^{*}	0.323
Content SS	0.185	0.164	0.180	0.159
Full-Time	0.185	0.099	0.110	0.059
APD			0.007^{**}	0.432
R^2	0.103		0.285	
F	1.213		3.463^{**}	
ΔR^2	0.103		0.183	
ΔF	1.213		13.303^{**}	

Note. N=59

Table 46 exhibits the full model, where, clear, established, school B, content SS, full-time, and APD (Model 2) are statistically significant in predicting collegiality, $R^2 = .228$, $F(6, 52) = 2.559$, $p = .030$; adjusted $R^2 = .139$. The addition of APD to the model leads to a statistically significant increase in R^2 of .149, $F(1, 52) = 10.040$, $p = .003$. The overall model accounts for 23% of the variation in in collegiality, with 15% attributed strictly to APD (8% attributed to teacher characteristics). Model 2 produces the following equation: Collegiality = .1 – (.238 x Clear) – (.027 x Established) + (.236 x School B) - (.022 x Content SS) - (.317 x Full-Time) + (.006 x APD); where a .72 SD increase in collegiality is expected from 120 hours (maximum participation) of APD.

Table 46.

Hierarchical Multiple Regression Analysis Predicting Collegiality Score from Clear Credential, Established at School Site, School, Content Area, Full-Time Teaching Status, and Academic PD Total Hours of Participation

	Collegiality			
	Model 1		Model 2	
	B	β	B	β
Constant	0.276		0.100	
Clear Credential	-0.275	-0.152	-0.238	-0.131
Established	0.035	0.027	-0.027	-0.021
School B	0.239	0.193	0.236	0.191
Content SS	-0.017	0.015	-0.022	-0.019
Full-Time	-0.247	-0.127	-0.317	-0.164
APD			0.006^{**}	0.319
R^2	0.079		0.228	
F	0.908		2.559^{*}	
ΔR^2	0.079		0.149	
ΔF	0.908		10.040^{**}	

Note. N=59

Table 47 illustrates the predictive power of the 5 covariates and APD on Leadership. Model 2 is statistically significant, $R^2 = .269$, $F(6, 52) = 3.19$, $p = .01$; adjusted $R^2 = .185$. The addition of APD to the model leads to a statistically significant increase in R^2 of .198, *F*(1, 52) = 14.089, *p* < .001. The overall model accounts for 27% of the variation in in leadership with 20% attributed strictly to APD (7% attributed to teacher characteristics). Model 2 produces the following equation: Leadership = -.477 – (.377 x Clear) + (.071 x Established) + (.292 x School B) - (.119 x Content SS) + (.246 x Full-Time) + (.010 x APD); where a 1 SD increase in leadership is expected from 100 hours (middle of the high participation group) of APD.

Table 47.

Hierarchical Multiple Regression Analysis Predicting Leadership Score from Clear Credential, Established at School Site, School, Content Area, Full-Time Teaching Status, and Academic PD Total Hours of Participation

	Leadership			
	Model 1		Model 2	
	B	β	B	β
Constant	-0.187		-0.477	
Clear Credential	-0.439	-0.170	-0.377	-0.146
Established	0.174	0.094	0.071	0.038
School B	0.298	0.168	0.292	0.165
Content SS	-0.112	-0.067	-0.119	-0.072
Full-Time	0.362	0.131	0.246	0.089
APD			0.010^{***}	0.450
R^2	0.071		0.269	
F	0.810		3.190^{*}	
ΔR^2	0.071		0.198	
ΔF	0.810		14.089^{***}	

Note. N=59

The last hierarchical multiple regression model conducted with the teacher survey constructs is for reflective practice. Model 2 (Table 48) is statistically significant, R^2 = .391, *F*(6, 52) = 5.560, *p* < .001; adjusted R^2 = .321. The addition of APD to the model leads to a statistically significant increase in R^2 of .308, *F*(1, 52) = 26.296, *p* < .001. The overall model accounts for 39% of the variation in in reflective practice with 31% attributed strictly to APD

(8% attributed to teacher characteristics). Model 2 produces this equation: Reflective Practice (RP) = -.474 – (.430 x Clear) - (.304 x Established) + (.438 x School B) - (.006 x Content SS) + (.480 x Full-Time) + (.013 x APD); where a 1.56 SD increase in reflective practice is expected from 120 hours of APD.

Table 48.

Hierarchical Multiple Regression Analysis Predicting Reflective Practice Score from Clear Credential, Established at School Site, School, Content Area, Full-Time Teaching Status, and Academic PD Total Hours of Participation

	Reflective Practice			
	Model 1		Model 2	
	B	β	B	β
Constant	-0.114		-0.474	
Clear Credential	-0.507	-0.196	-0.430	-0.167
Established	-0.176	-0.095	-0.304	-0.165
School B	0.445	0.252	0.438	0.248
Content SS	0.003	0.002	-0.006	-0.004
Full-Time	0.624	0.226	0.480	0.174
APD			0.013^{***}	0.561
R^2	0.083		0.391	
F	0.956		5.560^{***}	
ΔR^2	0.083		0.308	
ΔF	0.956		26.296^{***}	

Note. N=59

Because APD is not strongly correlated with either ongoing learning or collaboration, multiple regression models containing only the 5 covariates of interest are produced for these two teacher effectiveness constructs. Table 49 displays the results for the regression model predicting ongoing learning. Clear credential, established, school B, content SS, and full-time statistically significantly predict ongoing learning, $R^2 = .291$, $F(5, 53) = 4.338$, $p = .002$; adjusted $R^2 = .223$. This set of teacher characteristics accounts for approximately 29% of the variation in ongoing learning, with clear credential exerting the greatest effect. The following equation is produced: Ongoing Learning = 1.184 – (2.402 x Clear) + (.178 x Established) + (.166 x School B) – (.101 x Content SS) + (.427 x Full-Time). A teacher with these characteristics has a predicted score of -.55 on ongoing learning (just below the mean of 0).

Table 49.
Multiple Regression Analysis Predicting Ongoing Learning Score from Clear Credential, Established at School Site, School, Content Area, and Full-Time Teaching Status

	B	SE B	β
Constant	1.184	0.655	
Clear Credential	-2.402^{***}	0.548	-0.567
Established	0.178	0.407	0.059
School B	0.166	0.409	0.057
Content SS	-0.101	0.330	-0.037
Full-Time	0.427	0.561	0.094
R^2	0.290		
F	4.338^{**}		

Note. N=59

Table 50 displays the results for the multiple linear regression model predicting collaboration. Clear credential, established, school B, content SS, and full-time statistically significantly predict collaboration, $R^2 = .221$, $F(5, 53) = 3.03$, $p = .018$; adjusted $R^2 = .147$. This set of teacher characteristics accounts for approximately 22% of the variation in collaboration, with school B exerting the greatest effect. The following equation is produced: Collaboration = -.644 – (.171 x Clear) - (.103 x Established) + (1.032 x School B) + (.167 x Content SS) - (.017 x Full-Time). A teacher with these characteristics has a predicted score of .26 on collaboration (just above the mean of 0).

Table 50.
Multiple Regression Analysis Predicting Collaboration Score from Clear Credential, Established at School Site, School, Content Area, and Full-Time Teaching Status

	B	SE B	β
Constant	-0.644	0.487	
Clear Credential	-0.171	0.407	-0.057
Established	-0.103	0.303	-0.048
School B	1.032**	0.304	0.503
Content SS	0.167	0.245	0.086
Full-Time	-0.017	0.416	-0.005
R^2	0.221		
F	3.003*		

Note. N=59

A hierarchical multiple regression analysis is also run using the 5 covariates and APD to predict overall strategy use (expert assessment). Table 51 shows that Model 2 is statistically significant, $R^2 = .479$, $F(6, 42) = 6.426$, $p < .001$; adjusted $R^2 = .404$. The addition of APD to the model leads to a statistically significant increase in R^2 of .321, $F(1, 42) = 25.837$, $p < .001$. The overall model accounts for 48% of the variation in in overall strategy use with 32% attributed strictly to APD (16% attributed to teacher characteristics). Model 2 produces the following equation: Overall Strategy Use = 1.916 – (.110 x Clear) - (.214 x Established) + (.2 x School B) - (.536 x Content SS) + (.317 x Full-Time) + (.014 x APD); where a 1 SD increase in overall strategy use is expected from 72 hours (beginning of the high participation group) of APD.

For the classroom observation data, hierarchical multiple regression is conducted using total strategies to predict classroom environment (Table 52) and instruction (Table 53) while controlling for teacher characteristics. For classroom environment, Model 2 is statistically significant, $R^2 = .693$, $F(6, 23) = 8.663$, $p < .001$; adjusted $R^2 = .613$. The addition of total

Table 51.
Hierarchical Multiple Regression Analysis Predicting Overall Strategy Use from Clear Credential, Established at School Site, School, Content Area, Full-Time Teaching Status, and Academic PD Total Hours of Participation

	Overall Strategy Use			
	Model 1		Model 2	
	B	β	B	β
Constant	2.360^{***}		1.916^{***}	
Clear Credential	-0.076	-0.028	-0.110	-0.041
Established	-0.108	-0.050	-0.214	-0.099
School B	0.175	0.086	0.200	0.098
Content SS	-0.626^{*}	-0.327	-0.536^{*}	-0.280
Full-Time	0.500	0.172	0.317	0.109
APD			0.014	0.575
R^2	0.158		0.479	
F	1.612		6.426^{***}	
ΔR^2	0.158		0.321	
ΔF	1.612		25.837^{***}	

Note. N=49

strategies to the model leads to a statistically significant increase in R^2 of .556, $F(1, 23)$ = 41.706, $p < .001$. The overall model accounts for 69% of the variation in classroom environment with 56% attributed strictly to total strategies (13% attributed to teacher characteristics). Model 2 produces the following equation: Classroom Environment = .795 – (.108 x Clear) + (.071 x Established) + (.383 x School B) + (.217 x Content SS) + (.483 x Full-Time) + (.18 x Total Strategies); where a 1-point increase in classroom environment score (range of 1-4) is expected from the employment of 6 distinct instructional strategies per class.

Hierarchical multiple regression is also conducted using total strategies to predict instruction. The 5 covariates and total strategies are statistically significant predictors of instruction (Table 53, Model 2), R^2 = .801, $F(6, 23)$ = 15.437, $p < .001$; adjusted R^2 = .749. The addition of total strategies to the model leads to a statistically significant increase in R^2 of .749, $F(1, 23)$ = 80.513, $p < .001$. The overall model accounts for 80% of the variation in in

Table 52.
Hierarchical Multiple Regression Analysis Predicting Classroom Environment Score from Clear Credential, Established at School Site, School, Content Area, Full-Time Teaching Status, and Total Strategy Use

	Classroom Environment			
	Model 1		Model 2	
	B	β	B	β
Constant	1.690*		0.795	
Clear Credential	-0.263	-0.114	-0.108	-0.047
Established	0.057	0.035	0.071	0.044
School B	0.232	0.159	0.383	0.263
Content SS	0.392	0.282	0.217	0.156
Full-Time	0.803	0.350	0.483	0.211
Total Strategy Use			0.180	0.782
R^2	0.137		0.693	
F	0.762		8.663***	
ΔR^2	0.137		0.556	
ΔF	0.762		41.706***	

Note. N=30

Table 53.
Hierarchical Multiple Regression Analysis Predicting Instruction from Clear Credential, Established at School Site, School, Content Area, Full-Time Teaching Status, and Total Strategy Use

	Instruction			
	Model 1		Model 2	
	B	β	B	β
Constant	1.864*		0.953*	
Clear Credential	-0.488	-0.234	-0.331	-0.159
Established	-0.041	-0.028	-0.027	-0.018
School B	0.383	0.289	0.538**	0.405
Content SS	0.207	0.164	0.029	0.023
Full-Time	0.572	0.274	0.247	0.118
Total Strategy Use			0.183	0.875
R^2	0.105		0.801	
F	0.561		15.437***	
ΔR^2	0.105		0.696	
ΔF	0.561		80.513***	

Note. N=30

instruction with 75% attributed strictly to total strategies (5% attributed to teacher characteristics). Model 2 produces the following equation: Instruction = .953 – (.331 x Clear) - (.027 x Established) + (.538 x School B) + (.029 x Content SS) + (.183 x Full-Time) + (.18 x Total Strategies); where a 1-point increase in instruction score (range of 1-4) is expected from the employment of 6 distinct instructional strategies during a class period. Similarly, a 2-point increase is expected from the use of 12 distinct instructional strategies during a single class period.

Summary across Measures

For the most part, measures of teacher effectiveness are positively correlated with one another, largely supporting the claim that these are all measures of teacher effectiveness. This remains true within each instrument but does not always hold across instruments. For example, total strategy use, classroom environment, and instruction are highly positively correlated with each other but not with other constructs found on the teacher survey or expert assessment. The constructs contained within the teacher survey and expert assessment tend to be positively and often significantly correlated with one another and with APD participation. Findings seem to be triangulated across the teacher survey and expert assessment measures, but not necessarily to the classroom observation data. It appears that the classroom observation data provides unique information that is not found in or triangulated by the other data sources. This provides a great deal of support in deciding to keep classroom observations as part of a comprehensive teacher evaluation system.

In this sample, teacher characteristics, alone, do a relatively poor job of predicting scores on teacher effectiveness constructs. Table 54 summarizes the percent of the variance explained by each block of independent variables across the models (according to the order in which they are entered). The five teacher characteristics (clear credential, established at school site, school B, content SS, and full-time teaching status) predict between 5 to 29 percent of the variation in scores on dependent variables (instruction and ongoing learning, respectively). Only clear credential, school B, and content SS stand alone as independent variables that significantly predict scores on teacher effectiveness constructs. For only ongoing learning and collaboration are teacher characteristics solely significant in predicting teacher effectiveness.

Table 54.
Percent of Variance Explained by Covariates and APD or Total Strategies by Teacher Effectiveness Construct

	Percent of Variance Explained					
	Covariates	APD	Total	Covariates	Total Strategies	Total
Instructional Efficacy	11	18	29	-	-	-
Collegiality	8	15	23	-	-	-
Leadership	7	20	27	-	-	-
Ongoing Learning	29	-	29	-	-	-
Reflective Practice	8	31	39	-	-	-
Collaboration	22	-	22	-	-	-
Overall Strategy Use	16	32	48	-	-	-
Classroom Environment	-	-	-	13	56	69
Instruction	-	-	-	5	75	80

The intervention ITQP (APD participation) does a better job of predicting scores on dependent variables housed within the teacher survey and expert assessment. Between 15 to 32 percent of the variation in collegiality and overall strategy use, respectively, is explained by the intervention. Combined with the amount of variation explained by teacher characteristics, a moderate level of variance is explained by the final models. The count of total strategies used within a class period is a very strong predictor (between 56% for classroom environment and

75% for instruction) of scores on teacher effectiveness constructs contained within the classroom observation protocol.

Teacher characteristics, academic PD participation, and total strategy use, taken as a whole, are often good predictors of scores on teacher effectiveness constructs. Total variation explained by the collection of independent variables accounts for 22 (collaboration) to 80 (instruction) percent of the variation in scores on teacher effectiveness construct across instruments. A much greater proportion of the variance in scores on teacher effectiveness constructs is explained in the classroom observation data. For the teacher survey and expert assessment, at least half of the variation in scores remains unexplained.

Research Question 3

The following research question and sub-questions are addressed in this section: (3.) In what ways are individual teachers depicted similarly and differently across different measures of teacher effectiveness? (a.) In what ways do these patterns hold across teacher effectiveness constructs? (b.) In what ways do these patterns hold across measures?

Understanding emergent patterns for individual teachers across different measures of teacher effectiveness requires having a standardized way to compare scores on constructs. The standard score (z-score) is a useful statistics because it allows for comparison of scores from different normal distributions. Expressed as standard deviation from their means (mean 0, SD 1), a zed score and the standard normal distribution table can help discern the probability that a score will fall above or below a specific z-score. By using z-scores, I am uniformly able to create percentile rankings for teachers across constructs. Specifically, I am interested to see if teachers scoring above or below the mean, consistently score that way across constructs and measures. For this analysis I recorded quintile placement for each case (max 69) across all constructs (max 21 constructs and max 3 composite constructs - bold) over time. I also note (in grey) the cases in which teachers are scoring in the 90th or 10th percentile for a given construct.

One of the difficulties in using z-scores is that they are calculated based on a normal distribution, which is not always the case for this data. Multiple transformations were attempted to ameliorate the slight to moderate non-normality for some constructs, but due to group differences, no one transformation is able to transform the data to a perfectly normal distribution. Although this creates an imperfect estimate for quintile placement, the normality violations are inherently part of the data, which is not necessarily a bad thing. Keeping this in mind I primarily focus on teachers scoring above and below the mean, with particular attention paid to the teachers at the ends of the spectrum, where one expects to see the greatest differences. With this focus on teachers who have scored well above or below the mean, across multiple constructs, the effect of any error that is present due to non-normality is minimized.

In order to understand the ways in which quintile placement holds across constructs within a measure, 3 pattern analyses are conducted: classroom observation data; expert assessment data; and teacher survey data. Additionally, to understand the ways in which quintile placement holds across measures, 2 pattern analyses are conducted: teacher survey data and expert assessment data from year 2; teacher survey data, expert assessment data, and classroom observation data from year 3. Through close study, both quantitatively and qualitatively, of quintile placement for each case across constructs and measures, a more complete picture emerges about both, teacher effectiveness (and overall teacher quality) and the measures used to generate those estimates.

Classroom Observation Data Pattern Analysis

Close inspection of the classroom observation data reveals three primary patterns that characterize teacher scores across constructs. Table 55 lists quintile placement for a selection of teachers across constructs (composite constructs are bolded, 10th and 90th percentile placements are indicated in grey). The first primary pattern is one where teachers score primarily (at least 70% of the time or more) in the lower quartiles (for the 10 constructs) (12/30 cases – 40%). These teachers consistently rank below average across constructs (see cases 44, 24, 33, 36, and 61). Additionally, the teachers largely ranking in the 10th percentile also primarily place in the lowest quartile for the instances in which they are not in the 10th percentile (shown in grey, cases 44 and 24).

Table 55.
Example Quintile Placement Patterns from Classroom Observation Data

Case	Total Strategy Usage	**Classroom Environment**	Respect and Rapport	Culture for Learning	Classroom Procedures	Student Behavior	**Instruction**	Communicating with Students	Questioning and Discussion	Structures to Engage Students	Assessment for Learning	Flexibility and Responsiveness
44	1	1	1	1	1	1	1	1	1	1	1	1
24	2	1	1	1	1	1	1	1	1	1	1	1
33	1	2	2	1	1	2	1	1	2	1	2	1
36	1	2	2	2	2	2	2	2	2	2	2	2
61	2	2	3	2	2	3	2	2	1	2	2	4
19	3	4	5	4	4	2	2	1	2	3	2	2
51	3	3	3	4	4	2	3	3	2	3	3	2
20	4	3	3	4	4	2	3	3	3	4	2	4
52	2	4	4	5	4	4	4	4	4	4	4	4
14	5	5	5	5	5	4	5	5	5	5	4	4
47	5	5	5	5	5	5	5	5	5	5	5	5
55	5	5	5	5	5	5	5	5	5	5	5	5

The second primary pattern is one where teachers score primarily (at least 70% of the time or more) in the upper quartiles (for the 10 constructs) (11/30 cases – 37%). These teachers consistently rank above average across constructs (see cases 52, 14, 47, and 55). The teachers largely ranking in the 90th percentile also primarily place in the highest quartile for the instances in which they are not in the 90th percentile (shown in grey, cases 47 and 55). For teachers that primarily rank above average, seldom is there an instance of a low quartile placement (or visa versa) (see case 52).

The third primary pattern identified is for teachers consistently ranking around the mean, with some distribution across all three quintiles (7/30 cases – 23%) (see cases 19, 51, and 20). For these teachers, there is no clear consistency in scores across constructs; they appear to be strong in some areas and weaker in others. These are most likely developing teachers split across quintiles reflecting their evolving effectiveness.

Close examination of the classroom observation data reveals great consistency in scores across constructs, especially for the classroom environment and instruction constructs. By very large numbers, teachers either rank consistently below average or above average (23/30 cases – 77%). For this measure of teacher effectiveness, creating teacher effectiveness composite variables is suitable (Strategy Use, Classroom Environment, and Instruction), if desired.

Expert Assessment Data Pattern Analysis

There are 51 teachers for which expert assessment data is available. For many of those teachers (48), Spring Y2 and Spring Y3 data is available (totaling 99 distinct entries). I begin by looking at patterns in teacher scores across the 5 distinct instructional strategy categories for the 99 entries. Table 56 contains cases that demonstrate the emergent patterns for the expert assessment data across strategy type (composite construct in bold, upper and lower 10th percentile in grey). Four primary patterns emerge and 6 corresponding sub-patterns emerge.

Table 56.

Example Quintile Placement Patterns from Expert Assessment Data, Spring 2 and Spring 3

		Strategy Use					
Spring		**Overall**	Reading	Writing	Inquiry	Collaborative	Other
3	57	1	1	1	1	1	1
2	17	1	1	1	2	1	2
3	51	1	2	2	1	1	2
3	53	2	1	2	3	2	2
3	34	1	4	2	1	1	1
2	34	2	3	3	2	1	2
2	26	3	3	3	3	3	2
2	6	3	3	3	3	3	4
2	40	3	1	3	3	5	2
3	68	3	4	4	3	2	2
3	44	4	2	2	5	5	5
2	29	4	5	5	3	3	4
3	58	4	4	5	4	4	4
3	12	4	4	4	4	4	4
3	60	4	4	4	4	2	4
2	14	5	5	5	5	5	2
3	27	5	4	5	5	5	5
3	35	5	5	5	5	5	5

The first primarily pattern is for 40% of teachers who score consistently low in their quintile placement (see cases 57, 17, 51, 53, 34, and 24). These teachers place below average 60% of the time or more. Within this set of cases, there are 2 secondary patterns that help describe the distribution of the data. The first is for 5 teachers (13%) that have 60/40 split scores below average and average, respectively (see case 24). Although these teachers rank primarily in the lower quintiles, they have 40% of their scores close to the mean. The second sub-pattern is very rare (2 cases – 5%) and is characterized by teachers that place in the bottom quintiles 80%

of the time but have one constructs for which they score in an upper quartile (see case 34). The remaining 83% of the cases (33) fit the primary pattern, scoring in the 1st or 2nd quintile across 4 or 5 instructional strategy constructs. Usually, the teachers that fall into the lowest 10th percentile also score in the bottom quintile for other constructs (see case 57).

The second primary pattern is one that fits 36% of the cases and is characterized by teachers that score consistently in the top two quintiles across the 5 constructs (see cases 29, 58, 12, 60, 14, 27, and 35). These teachers place above average at least 60% of the time or more. Within this group, there are 2 sub-patterns worth noting. The first fits 4 cases (11%) and is characterized by a teacher scoring in the top two quintiles 60% of the time and the average quintile 40% of the time (see case 29). These teachers although above average, score near the mean 2/5 times. The second sub-pattern fits 8 cases (22%) and is characterized by teachers scoring in the top two quintiles 80% of the time with one construct for which they have a below average quintile placement (see cases 60 and 14). These teachers are consistently ranked above average except for one random low ranking. The remaining 67% of the teachers that fit this primary pattern score at least 80% of the time in the 4th or 5th quintile. Usually the teachers that score in the 90th percentile also fall within the top quintile for other constructs (see cases 27 and 35).

The third primary pattern covers 17% of the cases and is characterized by split quintile placements across constructs (see cases 68 and 44). These teachers fit two sub-patterns and are either spread out across the top, middle, and bottom quintiles (see case 68) (65% of split cases), or polarized between the highest and lowest quintiles 60/40 or 40/60 percent of the time (see case 44) (35% of split cases). Although these teachers may slightly lean above or below average, the variation in construct quintile placement makes their overall strategy (composite) score less reliable.

The last primary pattern fits 6% of the cases and is characterized by teachers that consistently place in the 3rd quintile across the 5 instructional strategy constructs (see cases 26 and 6). In three cases these teachers lean slightly to the lower end (see case 26) and in 3 cases these teachers lean slightly to the higher end (see case 6). Overall strategy scores are reliable for this group of teachers and overwhelmingly for the sample of teachers as a whole (83%). Using a composite score (overall strategy use) is reasonable for the expert assessment data.

Secondly, I look for patterns over time within the expert assessment data (Table 57). There are 48 teachers for which expert assessment data is available in Spring Y2 and Spring Y3. The groups identified above tend to behave differently over time (consistently low, consistently high, split, consistently average). For teachers that consistently score in the 1st and 2nd quintiles at Spring 2, they tend to have similar quintile placement at Spring 3 (see cases 54 and 13) or get worse (see cases 13 and 23). For these same teachers, the teachers in the lowest 10th percentile (in grey) remain extremely low into Spring 3 (see case 23). The teachers consistently scoring in the 4th and 5th quintiles at Spring 2 tend to have similar quintile placement at Spring 3 (see case 64) or get better (see cases 31 and 64). Similarly, the teachers in the 90th percentile (in grey) (Spring 2) remain extremely high into Spring 3 (see case 47).

The teachers scoring in the split and average quintile placement patterns behave erratically from one year to the next. Some of these teachers see an increase in quintile placement (see cases 4 and 19), or get worse (see cases 26 and 40) and generally show great fluctuations in their quintile scores (see cases 4, 19, 26, and 40). Overall, teachers tend to behave consistently over time in their below average, average, and above average quintile placements.

Table 57.

Example Quintile Placement Patterns over Time for Expert Assessment

	Strategy Use Spring 2						Strategy Use Spring 3					
Case	**Overall**	Reading	Writing	Inquiry	Collab.	Other	**Overall**	Reading	Writing	Inquiry	Collab.	Other
54	1	1	1	2	1	2	1	2	2	1	1	2
13	1	1	1	2	1	2	1	1	1	1	1	1
23	1	1	1	2	1	2	1	1	1	1	1	1
26	3	3	3	3	3	2	1	4	2	1	1	1
40	3	1	3	3	5	2	1	1	2	3	1	1
19	2	3	1	2	1	4	3	4	2	3	2	4
4	3	3	3	2	5	2	4	5	4	3	5	2
31	5	3	3	5	5	5	5	4	4	4	5	5
64	5	5	5	5	5	5	5	5	5	5	5	5
47	5	5	5	5	5	5	5	5	5	5	5	5

Typically, the low remain the same or get lower and the high remain high or get higher (more variation is found within the teachers scoring in the middle.

Lastly, I look strictly at the scores on overall strategy use over time. Fluctuation in quintile placement, with patterns varying by content area, is revealed. In science, 50% of teachers drop in quintile placement, while 40% remain constant and 10% increase. In social studies, 14% of teachers drop in quintile placement, which 43% remain constant and 43%increase. For reasons unknown, social studies teachers in this sample seem to be scoring higher than science teachers in Y3. In 1 case (science), a teacher moved from an above average score to a below average score and in 3 cases (science) moved from an average score to a below average score. In 5 cases (2 science, 3 social studies) a teacher moved from a below average score to an average score, in 2 (social studies) cases a teacher moved from a below average score to an above average score, and in 4 cases (social studies) a teacher moved from an average score to an above average score. Overall, only 4 teachers moved significantly downward over time in overall strategy use (all science), while 11 teachers significantly moved upwards over time in overall strategy use (2 science and 9 social studies), again pointing to the relative consistency of below average, average, and above average ratings. For the expert assessment measure of teacher effectiveness, creating a composite variable (overall strategy use) is reasonable.

Teacher Survey Data Pattern Analysis

Teacher survey data is available for three time points: Spring Y1, Spring Y2, and Spring Y3. Data is available for a total of 69 teachers, many of which remained in the study longitudinally (for a total of 168 total cases). 5 primary patterns are identified (see Table 58 for examples) within the teacher survey data. Teachers that have scores across the bottom two quintiles, middle quintile, and top two quintiles characterize the first, and most populous pattern. These teachers are part of the split quintile pattern and comprise 45% (75) of all cases (see cases 24, 27, 35, 42, 46, and 62). Within the split pattern, teachers may fall into one of three categories; evenly split, low split, and high split. Teachers that are evenly split place twice in the below average quintiles, twice in the average quintile, and twice in the above average quintiles, comprising 12% of the total (9/75 cases) (see cases 35 and 42). Teachers that are low split comprise 41% of the total (31/75 cases) and place in the bottom two quintiles at least 50% of the time with scores on the remaining constructs being spread across the mid and upper quintiles (see cases 24 and 62). 35 teachers (47%) remain in the high split group and place in the upper two quintiles at least 50% of the time with remaining constructs scores falling in the mid and lower quintiles (see cases 27 and 46). The split group teachers seem to bounce all over in their scores, garnering no definitively clear estimate of teacher effectiveness.

Table 58.
Example Quintile Placement Patterns for Teacher Survey, Spring 1, 2, and 3

Spring	Case	Instructional Efficacy	Collegiality	Leadership	Ongoing Learning	Collaboration	Reflective Practice
3	25	1	1	2	1	1	1
1	7	1	1	2	2	1	-
2	37	1	1	1	4	2	2
3	14	4	2	1	1	1	1
1	30	1	3	3	1	1	-
2	2	2	3	2	2	1	2
3	68	3	1	2	3	3	3
1	11	3	3	3	3	5	-
3	60	1	4	1	1	4	1
3	54	4	4	2	1	1	1
2	9	4	1	4	5	1	2
2	56	5	4	2	1	5	2
1	43	2	4	1	5	4	-
3	19	5	4	5	4	1	1
1	62	4	3	2	3	2	-
2	24	1	4	4	2	3	2
2	35	3	4	4	3	2	2
2	42	3	4	2	5	3	2
1	46	5	5	3	1	3	-
3	27	5	5	3	1	3	4
2	20	5	1	5	5	4	5
1	67	5	5	5	2	5	-
1	40	4	3	3	5	5	-
3	3	5	4	5	5	3	5
2	67	5	4	5	4	5	5
1	59	5	4	5	5	5	-

The next three primary patterns each comprise 15% (25-26/168 cases) of the teacher survey data. There are 26 teachers for which quintile placement is polarized across below average and above average scores. These teachers never place in the average quintile and are either spread 50/50 across the bottom and top two quintiles (15% of the polarized cases) (see cases 9 and 56), fall into the 1st and 2nd quintiles 67% of the time and the 4th and 5th quintiles 33% of the time (35% of the polarized cases) (see cases 54 and 60), or, fall into the 4th and 5th quintiles 67% of the time and the 1st and 2nd quintiles 33% of the time (50% of the polarized cases) (see cases 19 and 43). For the polarized teachers, it is also difficult to discern a clear estimate of teacher effectiveness, as they seem to be very strong in some areas and very weak in others.

Some teachers (15% - 25/168 cases) do score consistently low across teacher survey constructs. These teachers rarely have above average quintile placements and are heavily centered on the 1st and 2nd quintiles. 68% (17/25) of these cases score in the bottom two quintiles 60% of the time or more with no upper quintile placement (see cases 2 and 30). 3 of those teachers have pure scores – 100% of the time in the lowest quartiles (see cases 7 and 25). 32% of the low teachers (8/25) place in the 1st and 2nd quintiles 80% of the time and have one random upper quintile placement (see cases 14 and 37). The low teachers largely show consistency in their lower quintile placement across constructs.

Additionally, some teachers (15% - 26/168 cases) score consistently high across teacher survey constructs. These teachers rarely have below average quintile placements and are heavily centered on the 4th and 5th quintiles. 51% of these teachers (13/26) never have below average quintile placements and score in the 4th and 5th quintiles at least 60% of the time (see cases 3 and 40). For 3 of these cases, teachers have pure (100%) upper quintile placements (see cases 59 and 67Y2). 50% of the high teachers score in the 4th and 5th quintiles 80% of the time and have one random lower quintile placement (see cases 20 and 67Y1). The high teachers largely show consistency in their upper quintile placement across constructs.

The last pattern identified makes up 10% of the teacher survey cases. 16 teachers are heavily weighted in 3rd quintile placement across constructs (see cases 11 and 68). These average scoring teachers may slightly lean high or low but do not clearly differentiate themselves from the mean.

Teachers falling into any of the above patterns may also score in the 10th or 90th percentile (shown in grey) for one or more constructs. There does not seem to be much of a discernable pattern for the most extreme quintile placements. Generally, teachers have 1-3 (4 maximum) constructs in which they place in the lower and/or upper 10 quartiles (cases 2, 7, 9, 11, 14, 19, 20, 24, 25, 27, 30, 37, 42, 43, 46, 54, 60, and 67). Sometimes, teachers have both a construct score in the 10th and 90th percentile in any given year (see cases 9 and 20). There is also little consistency across years for these placements, even going from the highest placement one year to the lowest placement in another year (see case 51 in Table 59). The low consistency across years questions the reliability of the instrument. Creating a composite variable for the teacher survey data is not advisable.

Looking at teacher quintile placement over time yields additional information about teacher survey effectiveness measures. As a whole, teachers show great variations in their quintile placements from year to year. This makes it hard to strictly discern patterns amongst teachers because what appears to be a pattern between two years may not hold for a third year. With close examination, roughly 3 major patterns are discernable in the teacher survey data: teachers that approximately stay the same; teachers that show significant changes in quintile placement; and teachers that show erratic fluctuations in quintile placement. Table 59 displays examples of quintile placement patterns over time.

Table 59.

Example Quintile Patterns over Time, Teacher Survey

	Spring 1					Spring 2						Spring 3					
Case	Instructional Efficacy	Collegiality	Leadership	Ongoing Learning	Collaboration	Instructional Efficacy	Collegiality	Leadership	Ongoing Learning	Collaboration	Reflective Practice	Instructional Efficacy	Collegiality	Leadership	Ongoing Learning	Collaboration	Reflective Practice
13	-	-	-	-	-	1	1	1	1	1	2	1	1	1	1	1	1
2	1	3	2	2	2	2	3	2	2	1	2	2	1	2	2	1	1
25	1	1	2	3	3	1	2	3	3	1	2	1	1	2	1	1	1
1	-	-	-	-	-	1	1	1	1	4	2	1	3	5	5	3	1
30	1	3	3	1	1	5	5	3	1	2	2	-	-	-	-	-	-
66	2	2	5	2	1	3	4	5	2	3	2	-	-	-	-	-	-
8	1	4	5	5	5	2	2	3	1	4	2	-	-	-	-	-	-
59	5	4	5	5	5	4	2	3	3	4	2	4	2	3	4	5	4
14	-	-	-	-	-	4	4	4	2	4	2	4	2	1	1	1	1
17	-	-	-	-	-	1	4	4	4	5	2	1	3	2	2	4	1
57	-	-	-	-	-	4	4	4	1	4	4	3	3	2	1	3	1
51	5	5	5	4	4	5	5	5	3	5	4	5	1	4	2	5	4
45	5	5	5	1	5	5	5	4	3	4	2	5	5	5	2	5	5
47	3	2	5	4	4	3	4	5	1	4	3	5	4	5	2	4	5
67	5	5	5	2	5	5	4	5	4	5	5	4	4	5	3	5	5

The teachers that stay approximately the same fall into two categories: those that consistently place in the lower quintiles and those that consistently place in the higher quintiles. Cases 2, 12, and 25 consistently score in the 1st and 2nd quintiles with some placement into the 3rd quintile, over the three years. There is some variation in placement between years but the teachers primarily remain in the below average quintiles. Some of those teachers also place in the 10th percentile (in grey), which sometimes holds constant across years (see case 13, case 2 Spring 2 to Spring 3) and sometimes doesn't (see case 2 and 25 Spring 1 to Spring 2). Cases 45, 47, and 67 consistently score in the 4th and 5th percentile with occasional scores in the lower quintiles. Again there are slight variations across years and some matching 90th percentile (in grey) rankings (see cases 47 and 67).

A group of teachers exhibit either upward or downward movement over time in quintile placement. Cases 1, 30, and 66 all display increases in their quintile movement, in some cases moving from the 1st to the 5th quintile within a year's time (see cases 1 and 30), and in other cases moving from the bottom 10 percent for some constructs to the top 90th percentile for other constructs (see case 1). Cases 17, 51, and 57 exhibit a decrease in quintile placement over time. Case 51 largely stays the same between Spring 1 and 2 and then sees a significant decline in quintile scores at Spring 3, even moving from the 90th percentile in collegiality to the 10th percentile in the same construct one year later. Overall, there is significant movement in construct quintile scores over time signifying either growth/decline or instability in the measure itself.

Lastly, there is a group of teachers that show drastic changes in quintile placements from year to year, both positive and negative, making it hard to discern a true score on constructs. Cases 8, 14, and 59 display erratic changes in quintile scores over time. Sometimes these teachers place in above average quintiles one year and below average quintiles the next and often see scores increase/decrease 2-3 quintiles for multiple constructs between years. The lack of consistency over time for many teachers makes it difficult to make sense of the teacher survey effectiveness data. One must question the reliability of the instrument when individual scores fluctuate wildly. There are many contextual factors that can affect self-reported data (mood, timing of administration, program experiences, school site factors like funding and RIF notices, etc.), likely that is the case here, introducing threats to validity in any given year. When a measure is vulnerable to variation in scores (due to environmental or other factors) it looses its credibility for use in a comprehensive teacher evaluation system. Both the within year and across years quintile scores are so inconsistent that it would be hard to argue for inclusion in any legitimate teacher evaluation system

Teacher Survey and Expert Assessment Data Pattern Analysis

There are a total of 46 teachers in Year 2 for which both teacher survey and expert assessment data is available for an across measure pattern analysis. Given the high level of inconsistency within individual responses for the teacher survey data and the relatively high level of consistency within responses for the master teacher assessment data, comparisons across instruments is challenging. For this reason I am looking for rough matches in quintile placement patterns across measures. Table 60 shows that half of the time there is some level of matching between quintile patterns and half the time there is not.

Table 60.
Percent of the Time Quintile Placement Patterns Match between Teacher Survey and Expert Assessment, Spring 2

		% Match	
Instruments	n	Yes	No
Teacher Survey : Expert Assessment	46	50	50

Example quintile patterns are illustrated in Table 61. The group of teachers for which there is some level of matching quintile scores across measures largely fit into 4 groups. The first characterizes the teachers that consistently score in the lower quintiles (see cases 13 and 54).

For these teachers there is a definitive match in teacher effectiveness portrayal across instruments. For the other three groups with matching quintile patterns, teacher effectiveness is less definitive. There is a group of teachers that score largely in the upper quintiles but may have some conflicting scores in the teacher survey data (see cases 14, 45, and 51). There is also a group of teachers that exhibit some degree of match in quintile placement across measures but also a degree of uncertainty in the match (see cases 24 and 31). Lastly, there is a group of teachers that are consistently spread across quintile scores for each set of teacher effectiveness measures (see cases 42 and 53).

Table 61.
Example Quintile Patterns between Teacher Survey and Expert Assessment Data, Spring 2

Case	Instructional Efficacy	Collegiality	Leadership	Ongoing Learning	Reflective Practice	Collaboration	**Overall Strategy Use**	Reading	Writing	Inquiry	Collaborative	Other
13	1	1	1	1	2	1	1	1	1	2	1	2
54	3	3	2	1	2	2	1	1	1	2	1	2
53	1	4	3	2	2	4	2	3	1	2	1	4
42	3	4	2	5	2	3	4	3	3	2	3	5
24	1	4	4	2	2	3	1	1	1	2	1	2
31	4	5	2	5	5	1	5	3	3	5	5	5
17	1	4	4	4	2	5	1	1	1	2	1	2
35	3	4	4	3	2	2	5	5	5	5	5	5
67	5	4	5	4	5	5	1	1	1	2	1	2
51	5	5	5	3	4	5	2	3	1	2	1	2
47	3	4	5	1	3	4	5	5	5	5	5	5
14	4	4	4	2	2	4	5	5	5	5	5	2
46	4	4	4	2	4	4	5	5	5	5	5	5
45	5	5	4	3	2	4	5	5	5	5	5	4

For half of the teachers there is no match in quintile placement patterns across measures. Cases 17, 35, 51, and 67 show conflicting portrayals of teacher effectiveness between measures.

In some cases the mismatch is very pronounced (see cases 51 and 67) and in other less blatant (see cases 17 and 35). Regardless, there exists a significant degree of uncertainty of whether or not teachers are below average, average, or above average in the sample.

Inspection of the teachers that place in the lowest and highest 10th percentile (in grey) reveals low consistency between measures. Few cases (31, 42, and 47) score in the 90th percentile for at least one construct in each measure. In some cases, there is consistency in quintile placement but no carryover of the lower or upper bounds (see cases 13 and 46). For case 67 there is absolutely no carryover in 90th percentile ranking. The overall lack of consistency between measures complicates interpretation of teacher effectiveness quintile scores.

Teacher Survey, Expert Assessment, and Classroom Observation Data Pattern Analysis

There are 51 teachers in Year 3 that have at least two different data sources available for the multiple measure pattern analysis. Given the prevalence of different patterns within each instrument, looking across instruments poses challenges. In order to look across instruments, I look for rough matches in quintile placement. Because teachers are often split across quintiles for different constructs I look particularly at the way teachers are leaning in their scores. For example, a teacher that clearly places high in quintile scores for the classroom observation data and places split-high or average-high in teacher survey data is considered a match. Likewise a teacher that has polarized-low scores for the teacher survey data and consistent-low scores for expert assessment data is considered a match. If a teacher regularly has scores spread across quintiles but always leans low, that is again a match. When teachers are clearly low or high in quintile placements, the analysis is straightforward.

Teacher quintile pattern scores are largely arranged across two categories: matching (to some extent); and not matching at all. The majority of teachers (80% - 10/51 cases) match to some extent in quintile placement across measures of teacher effectiveness (2/2 instruments, 2/3 instruments, or 3/3 instruments). Table 62 shows example quintile patterns across measures at Year 3. Within the matching group we have: teachers that score consistently low (see cases 13 and 33); teachers that score consistently high (see cases 3 and 37); teachers that are consistently split across quintile scores (often leaning in one direction) (see cases 20 and 27); and teachers that have matching scores on 2/3 instruments but contradictory scores on a third (see cases 40, and 53 for two low matches and one contradictory high, and cases 14 and 45 for two high matches and one contradictory low).

For 20% of the teachers (11/51 cases), there is no match between quintiles scores across measures of teacher effectiveness (2/2 instruments or 3/3 instruments) (see cases 36, 51, and 67). These teachers seem to bounce all over the place in their quintile scores on teacher effectiveness constructs. They may have one high set of quintile scores and one low set of quintile scores (see case 67); one set of high quintile scores, one set of average quintile scores, and one set of low quintile scores (see case 36); or a wide range in quintile scores across constructs and measures (see case 51). For these teachers, it may be that they are truly developing; high in some areas and low in others. It may also be where the fabric of reliability is broken down. Regardless, deciphering any larger view of teacher quality from the individual measures of teacher effectiveness is impossible for these folks.

Table 62.

Example Quintile Placement Patters across Measures, Spring 3

	Teacher Survey						Expert Assessment						Classroom Observation											
Case	Instructional Efficacy	Collegiality	Leadership	Ongoing Learning	Reflective Practice	Collaboration	**Overall Strategy Use**	Reading	Writing	Inquiry	Collaborative	Other	Total Strategy Usage	**Classroom Environment**	Respect and Rapport	Culture for Learning	Classroom Procedures	Student Behavior	**Instruction**	Communicating with Students	Questioning and Discussion	Structures to Engage Students	Assessment for Learning	Flexibility and Responsiveness
13	1	1	1	1	1	1	1	1	1	1	1	1	-	-	-	-	-	-	-	-	-	-	-	-
33	1	1	2	3	1	1	-	-	-	-	-	-	1	2	2	1	1	2	1	1	2	1	2	1
53	3	3	4	1	4	3	2	1	2	3	2	2	1	1	1	1	1	1	1	2	1	1	1	1
40	3	1	2	1	2	5	1	1	2	3	1	1	5	5	5	5	4	5	5	4	5	4	5	5
67	4	4	5	3	5	5	1	1	1	1	1	1	-	-	-	-	-	-	-	-	-	-	-	-
51	5	1	4	2	4	5	1	2	2	1	1	2	3	3	3	4	4	2	3	3	2	3	3	2
36	3	4	3	2	3	3	5	5	5	5	5	5	1	2	2	2	2	2	2	2	2	2	2	2
20	3	2	3	3	4	3	3	2	2	4	4	2	4	3	3	4	4	2	3	3	3	4	2	4
27	5	5	3	1	4	3	5	4	5	5	5	5	2	4	5	3	4	4	4	4	4	4	5	4
14	4	2	1	1	1	1	4	4	4	5	5	2	5	5	5	5	5	4	5	5	5	5	4	4
45	5	5	5	2	5	5	5	5	5	3	5	4	2	1	2	1	1	1	1	1	1	1	1	2
3	5	4	5	5	5	3	4	4	4	3	5	5	5	3	3	4	3	3	4	4	4	4	4	4
47	5	4	5	2	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

Looking at teachers with the most extreme scores (in grey) yields important information. In the consistently low group there are multiple instances where that teacher also places in the 10th percentile for quartile scores (see cases 13 and 33). For the consistently low teachers there is no conflicting 90th percentile placement. Likewise, the consistently high teachers regularly place in the 90th percentile for their quartile scores but never in the 10th percentile (see cases 3and 47). Teachers in all other groups may have a mix of quartile scores in the 10th and 90th percentiles across measures (see cases 40, 67, 14, 45). When a teacher has major conflicting quartile placements like case 67, scoring in the top 90th percentile for 3/6 teacher survey effectiveness measures and in the bottom 10% for 2/5 expert assessment effectiveness measures, one has to grapple with ways to interpret the data. Such extreme differences based on the measure makes any teacher evaluation system centered on one measure highly suspect.

Table 63 shows the percentage of time quintile placement patterns roughly match across the different measures. The way teachers are depicted in the teacher survey data and expert assessment data is the most similar. 69% of the time quintile placement patterns align across these two instruments. The reverse situation is present for teacher survey data and classroom observation data. There is often (63% of the time) a mismatch in quintile placement patterns across these two instruments. About half of the time (54%) teachers are depicted similarly between the expert assessment and classroom observation measure of teacher effectiveness. Only a third of the time (32%) do teachers have matching quintile placements across all three measures of teacher effectiveness.

Table 63.
Percent of the Time Quintile Placement Patterns Match across Instruments, Spring 3

		% Match	
Instruments	n	Yes	No
Teacher Survey : Expert Assessment	49	69	31
Teacher Survey : Classroom Observation	30	37	63
Expert Assessment : Classroom Observation	28	54	46
Teacher Survey : Expert Assessment : Classroom Observation	28	32	68

Although there is some level of matching depictions of teacher effectiveness across measures, there still remains a significant portion of uncertainty in determining teacher effectiveness. As many as 20% of teachers (10/51 cases) never exhibit quintile patterns that match across instruments. Between any two instruments there is even greater mismatch – 31%

to 63%. It may be that teachers are naturally stronger and weaker in certain areas and that the mismatch between instruments is warranted. It could also mean that there is some degree of unreliability in the measures themselves.

When constructing holistic teacher evaluation systems, it is important to select measures that capture the complexity of the profession in valid, accurate, and reliable ways. It may in fact be the case that two measures do not align, hence strengthening the argument for a truly comprehensive system. However, in this case, due to the lack of consistency in scores within the teacher survey data by itself, I do not consider it a reliable instrument to measure teacher effectiveness, and therefore do not trust the different portrayal of effectiveness. The high alignment of the teacher survey with the expert assessment data also makes the expert assessment instrument suspect in my eyes. It is likely that these three instruments do a relatively good job identifying the lowest of the low and the highest of the high of the teachers in this sample. Where there is very strong alignment across measures, estimates are probably reliable.

It is also important to note that although breaking the teachers into 5 quartiles facilitates making comparisons across teachers in this sample (standardizes scores and creates a spread of below average, average, and above average scores), it does not help truly understanding teacher effectiveness placement as compared to other samples. As noted before, the teachers in this sample scored relatively low on many constructs (especially in classroom observations). An above average teacher in this sample may in fact be a below average teacher in another sample. Although this study does not definitively identify teacher effectiveness (classroom observations probably do the best job at this), it does do a good job of identifying the teachers that consistently score at the bottom or top of the sample. The lowest of the low in this sample likely compare with the lowest that you will find anywhere. To get an idea of the difference between the low and high scoring teachers across constructs and instruments, one can look at the bottom and top cut points for quintiles.

Table 64 shows the cut points for the 1st and 5th quintile by construct. For teacher survey constructs, the smallest mean differences between the 1st and 5th quintiles is a difference of 1.7 on a 5-point scale. The largest mean difference is 2.33 points on an 8-point scale. For the expert assessment data, the smallest and largest mean difference between the 1st and 5th quintiles is 1.0 (reading strategies) and 3.0 (collaborative and other strategies) on a 4-point scale, respectively. Lastly, the smallest and largest spread between the 1st and 5th quintiles for the classroom observation data is for structures to engage students in learning (.97) and questioning and discussion techniques (1.5), all on a 4-point scale. Across all of the constructs, the difference in top and bottom quintile cut points is significant. This indicates that there is in fact some differentiation between teachers at the school sites, for these measures of teacher effectiveness (it does not mean that those differences are aligned with groups based on ITQ APD participation).

Table 64.

Cut Points for 1st and 5th Quintiles by Construct and Composite Construct (Bold)

		Mean		
		1st Quintile	5th Quintile	Difference
Teacher Survey	Instructional Efficacy	3.33	5.00	1.70
	Collegiality	3.00	4.30	1.30
	Leadership	2.75	4.90	2.15
	Ongoing Learning	1.00	2.35	1.35
	Reflective Practice	1.00	3.33	2.33
	Collaboration	3.71	5.29	1.58
Expert Assessment	**Overall Strategy Use**	**1.48**	**3.40**	**1.92**
	Reading	2.00	3.00	1.00
	Writing	1.40	3.00	1.60
	Inquiry	1.00	3.60	2.60
	Collaborative	1.00	4.00	3.00
	Other	1.00	4.00	3.00
Classroom Observations	Total Strategy Use	3.10	8.40	5.30
	Classroom Environment	**1.94**	**3.31**	**1.37**
	Respect and Rapport	2.03	3.33	1.30
	Culture for Learning	1.90	3.25	1.35
	Classroom Procedures	1.67	3.13	1.46
	Student Behavior	1.87	3.33	1.46
	Instruction	**1.59**	**2.85**	**1.26**
	Communicating with Students	1.75	3.10	1.35
	Questioning and Discussion	1.50	3.00	1.50
	Structures to Engage Students	1.78	2.75	0.97
	Assessment for Learning	1.75	2.95	1.20
	Flexibility and Responsiveness	1.50	2.95	1.45

Summary across Pattern Analyses

Both the expert assessment and classroom observation data show high levels of consistency in quintile scores within a given year. The expert assessment data also displays a great deal of consistency in teacher quintile placement over time (unavailable for the classroom observation data). The estimates of teacher effectiveness produced with these two measures are largely reliable (for what they purport to measure). Producing composite teacher effectiveness scores for each of these measures is feasible. The teacher survey data on the other hand rarely shows consistency in quintile scores within a year or across years, raising serious questions about the reliability of the data. Only the very top and bottom teachers tend to have consistent quintile scores within a given year and across years. While the instrument may a do a reasonably good job at identifying the least and most effective teachers within the sample, it fails all the rest of the teachers. This eliminates the possibility of creating a composite score for the teacher survey effectiveness constructs, which further limits its use in a comprehensive teacher evaluation system.

Looking at how teachers are depicted across measures adds a great deal of uncertainty in discerning teacher effectiveness. To some extent there should be inconsistencies across constructs and measures. For a developing teacher it is perfectly understandable for scores in some areas to be higher or lower than in other areas. However, when this is largely the case something deeper is operating. There should not be extremely conflicting portrayals of teachers which is often the case in this sample when looking across measures. Undoubtedly, the data shows that the teacher survey data is unreliable. This may be due to the inherent aspects of the measure itself or its vulnerability to external factors. Regardless, there is little argument that can be made to support including the teacher survey in a comprehensive teacher evaluation system.

CHAPTER 5: Discussion

This study seeks to create an understanding of what three distinct measures of teacher effectiveness, namely, a teacher survey, expert assessment, and classroom observations, can contribute to an understanding of teacher quality. Through investigating the sensitivity level of each instrument (to detecting both between-group and within-group differences), the relationships between teacher effectiveness constructs (measures, and teacher characteristics), and the extent to which depictions of teachers vary across measures, I am able to make recommendations for comprehensive teacher evaluation systems (CTESs). Implications for CTESs are of immediate importance to those involved in educational research, educational program evaluation, Improving Teacher Quality programs, and ultimately teacher evaluation to better select measures for inclusion in subsequent systems, studies and evaluations.

This chapter begins with a brief summary and discussion of findings from the study. While this section contains a summary of results across research questions, a full discussion of findings for each research question is found in the prior chapter, at the end of each section. A discussion of implications for comprehensive teacher evaluation systems, implications for improving teacher quality professional development programs, study limitations, and possible directions for future research are also included in this chapter.

Summary of Results across Research Questions

Exploration of the measures used in this study provides both a unique understanding of measuring teacher effectiveness and the comparability of estimates gleaned from those measures. Through closely examining the measures of teacher effectiveness used in this study, a new understanding of each instrument's ability to detect differences (both between-groups and within groups), and the similarities (or differences) in the depictions of teachers across constructs and measures, is gained. In addition, a thorough understanding of the relationships between teacher effectiveness constructs and teacher characteristics is gained, including the predictive power of professional development (or total strategy use) on scores on teacher effectiveness constructs.

The three measures of teacher effectiveness used in this study are able to capture within-group differences (change over time) to varying extents. The teacher survey does a relatively poor job detecting change over time, only capturing changes in 2/6 teacher effectiveness constructs (ongoing learning and reflective practice) housed within the survey. The expert assessment does a good job capturing changes in overall strategy use within a year's time and across years. The classroom observation protocol only does a fair job detecting change over time in classroom environment and instruction (constructs and sub-constructs) for the high participation group. It is expected that given additional years of data, even greater sensitivity would be displayed for this measure.

Overall, few meaningful differences in scores on teacher effectiveness constructs are discovered between low, moderate, and high participation teachers (between-group differences). Across the board, the expert assessment proves itself to be the most sensitive to detecting both within-group and between-group differences. The teacher survey only captures significant group differences in year 3 of the study on instructional efficacy, leadership, and reflective practice. The classroom observation protocol fails at detecting any between-group differences in year 3 of the study.

For the most part, measures of teacher effectiveness are positively correlated with one another. This remains true within each instrument but does not always hold across instruments.

The constructs contained within the teacher survey and expert assessment tend to be positively and often significantly correlated with one another (and with APD participation) but not with constructs found on the classroom observation protocol. While classroom observation constructs are positively correlated with one another, the lack of alignment with other measures suggests unique information is captured by the data source, hinting to its importance for inclusion in a CTES.

In this sample, teacher characteristics, alone, do a relatively poor job of predicting scores on teacher effectiveness constructs (explaining between 5% to 29% of the variation in scores on dependent variables). Only credential, school, and content stand alone as independent variables significantly predictive of scores on teacher effectiveness constructs. For only ongoing learning and collaboration are teacher characteristics solely significant in predicting teacher effectiveness.

The intervention ITQP (APD participation) does a slightly better job predicting scores on dependent variables housed within the teacher survey and expert assessment (explaining 15% to 32% of the variation in construct scores). Combined with the amount of variation explained by teacher characteristics, a moderate level of variance is explained by the final models (at least half of the variation in scores remains unexplained). The count of total strategies used within a class period is a very strong predictor (between 56% for classroom environment and 75% for instruction) of scores on teacher effectiveness measures contained within the classroom observation protocol. A much greater proportion of the variance in scores on teacher effectiveness constructs is explained in the classroom observation data.

Both the expert assessment and classroom observation data show high levels of consistency in quintile scores within a given year. The expert assessment data also displays a great deal of consistency in teacher quintile placement over time (unavailable for the classroom observation data). The estimates of teacher effectiveness produced with these two measures are largely reliable (for what they purport to measure), alluding to the feasibility of producing composite teacher effectiveness scores for use in a CTES. The teacher survey data on the other hand rarely shows consistency in quintile scores within a year or across years, raising serious questions about the reliability of the data. Only the very top and bottom teachers tend to have consistent quintile scores within a given year and across years. This eliminates the possibility of creating a composite score for the teacher survey, which further limits its use in a CTES.

When looking at how teachers are depicted across measures some inconsistencies are expected, as teachers reasonably may be strong in some areas and weaker in others. However, comparisons in this study reveal more inconsistencies in the data than consistencies, raising questions about the reliability of some of the data, particularly, the teacher survey data, where extreme conflicting portrayals are evident both across years and across measures. The classroom observation data and the expert assessment data are aligned about half of the time, clearly indicating that they measure overlapping but different components of effectiveness. Unquestionably, classroom observations should be part of a holistic approach to teacher evaluation as they are the primary means for understanding what goes on in a classroom, and chiefly fulfill the formative role of teacher evaluation.

Implications for Comprehensive Teacher Evaluation Systems (CTESs)

Each of the measures used in this study assess different components of teacher effectiveness, potentially contributing something unique to a portrayal of teacher quality. This multi-measure investigation creates an understanding of each measure’s suitability for inclusion in a CTES. Through close examination of the teacher survey, expert assessment, and classroom observation data, important discoveries for CTESs are revealed. In this section, I address data collection considerations for large-scale implementation of CTES, suggested measure(s) for inclusion in a CTES, and recommendations for the use of classroom observations in CTESs.

Data Collection Considerations for Large-Scale Implementation of CTES

One important consideration in designing and implementing comprehensive teacher evaluation systems (CTESs) is the balance between the burden of data collection and the value of the data. Each of the measures used in this study requires a different level of resources (time, energy, money, materials) to collect and yields vastly different information. Additionally, the willingness to participate in the study varies widely based on the measure used, directly impacting the quality of the data and potential analyses. Table 65 shows the response rate for each measure of teacher effectiveness used in this study over time. In this section I assume that the information gained from each data source is valuable and strictly discuss how data collection challenges for each measure may influence a particular measure's use in a CTES. In particular I pay close attention to the feasibility and sustainability of replicating data collection on a large-scale.

For all three measures, response rates increase over time. This is likely due to changes in approaches to data collection, including procedures, incentives and strengthening of rapport. By the last administration of the teacher survey, an overwhelming 95% of teachers completed the survey. Assuming the data is valuable, the moderate level of resources (coordination, site visits, printing, incentives, data entry) required to get the teacher survey data and the stellar response make replication appealing. Given adequate resources, there is a moderate-high-level of feasibility of large-scale replication in sustainable ways.

Table 65.
Response Rate (%) by Measure over Time within Schools and Content Areas

	Teacher Survey			Expert Assessment			Classroom Observation	
	Y1T2	Y2T2	Y3T2	Y2T2	Y3T1	Y3T2	Y3T1	Y3T2
School A								
Science	100	100	90	42	50	50	20	40
Social Studies	78	100	92	92	100	100	17	50
Total	89	100	91	67	77	77	18	45
School B								
Science	80	90	100	81	85	85	30	35
Social Studies	70	90	95	81	85	85	40	50
Total	75	90	98	81	85	85	35	43
Study								
Science	87	94	97	67	73	73	27	37
Social Studies	72	94	94	85	91	91	31	50
Total	80	94	95	76	82	82	29	44

The expert assessment utilizes existing relationships between Professional Learning Partners (PLPs) (expert teachers) and teachers to obtain ratings on frequency of use of sets of instructional strategies. Given the position of instructional coach/professional learning partner/expert teacher is filled at a school site, a low level of resources (1-day PLP training, coordination) is required to obtain the data. Assuming the information is valuable and considering the reasonable response rate (82% at T3), one could imagine a high-level of large-scale replication in sustainable ways.

The response rate for the classroom observations is low – 44% (T2). Data collection for this measure is extremely time consuming and requires a high volume of resources (primarily in observer training and time for data collection and coding). In ITQP year 3, after much negotiation, only 15 teachers agreed to participation in 1 observation and 15 teachers agreed to participation in 2 observations. Being the most difficult to collect, large-scale sustainable replication has a low-level of feasibility. Assuming there is value to the data, inclusion of classroom observations in a CTES necessitates proper support structures and thoughtful implementation.

Suggested Measure(s) for Inclusion in CTESs

Each of the measures used in this study assess different components of teacher effectiveness, potentially contributing something unique to a portrayal of teacher quality. Ultimately, one has to decide on the "right" measures to include in a CTES. Through close examination of the teacher survey, expert assessment, and classroom observation data, I am able to assess each measure's suitability for inclusion in a CTES.

Findings from this study indicate that the teacher survey used is not suitable for inclusion in a CTES. The poor alignment of between-group differences in year 3 (instructional efficacy, leadership, and reflective practice) with within-group significant change over time (ongoing learning, and reflective practice) calls into question the impact on ITQP on participants and more importantly the validity of the instrument. It seems the instrument is overly prone to influence by contextual variables (as evidenced in the common decline in ongoing learning scores), indicating serious threats to internal validity. The lack of consistency in construct scores from one year to the next calls into question the reliability of the measure. The lack of consistency in construct scores within a year and across instruments in the same year also calls into question the reliability of the measure. Behavioral items (i.e. collaboration, leadership, reflective practice) housed within the survey may continue to be of use to professional development programs to gauge programmatic effects but not as a measure of teacher effectiveness. The inability to create a composite score from the measure also directly limits its use in CTES. Shortfalls of the teachers survey point to the need for alternative means to obtain a self-report of teacher effectiveness.

The expert assessment proves to be a reliable way to assess frequency of use of various sets of instructional strategies (by PLPs). The relative ease of data collection presents no problems for use in a CTES. The main question with the expert assessment measure of teacher effectiveness concerns the value of the information produced. Gauging frequency of use of instructional strategies can easily be done during classroom observations, which discounts the need for a separate instrument. The primary asset of the expert assessment is the involvement of PLPs in assessing teacher practice. Given PLP qualifications and their close work with teachers on a daily basis, PLPs are uniquely positioned to conduct expert evaluation. Moving forward with developing CTESs, one would be wise to consider including knowledgeable experts, such as PLPs, in the data collection process. One potential avenue for inclusion would be in the many facets of the classroom observation process.

Classroom observations prove to be both a valid and reliable way to assess teacher effectiveness. Providing an objective portrayal of teacher effectiveness by uniquely studying teaching practice, classroom observations are arguably the most important component of a comprehensive teacher evaluation system. The consistency in construct scores is very conducive to creating a composite score to assess overall teacher quality. Furthermore, classroom observations serve as the primary means to accomplish the formative roles of teacher evaluation and are treated as credible and fair when conducted by qualified observers. They also include a component of self-evaluation in the pre-formal-observation cycle procedures, giving teachers a much-needed voice. The biggest challenge to involving observation of teaching practice in a CTES is the feasibility of large-scale implementation. Using classroom observations in a CTES demands a high level of commitment from teachers, observers, schools, and the district. Considering the value of the data produced the effort required is manageable given adequate resources. Classroom observations, such as proposed in LAUSD, have the potential to deeply impact education when implemented with fidelity.

Recommendations for use of Classroom Observations in CTESs

The findings from this study strongly suggest the use of classroom observations in a comprehensive teacher evaluation system (CTES). One of the challenges (discussed more fully below in limitations) faced in this study is the overall low scores (and low variability in scores) on classroom environment and instruction components and elements in the classroom observation data. On a 4-point scale (highly effective, effective, developing, ineffective), teachers largely score in the developing to effective range (2-3) for this sample. Greater capacity to differentiate classroom environment and instruction practices could potentially increase the measure's sensitivity to detecting both between-group and within-group differences (as shown in teacher survey findings highlighting group differences for constructs with larger scales (7-point and 8-point)).

Assuming that classroom observations will continue to be used in CTESs, I recommend increasing the scale to a minimum of 5-points, similar to the Tennessee Educator Acceleration Model (significantly above expectations, above expectations, at expectations, below expectations, significantly below expectations) or the District of Columbia Public Schools IMPACT rating structure (highly effective, effective, developing, minimally effective, ineffective) (DC Public Schools website, 2009; TEAM TN website, 2014). A larger (more differentiated) scale could help capture greater variability in observable practice. This is especially important for largely ineffective teachers for which change (growth) over time is expected and required. Given the implications observation ratings may have (both formative and summative), it is vitally important that the sensitivity level of the measure's scale is conducive to achieving primary evaluation outcomes.

Additionally, systems and structures must be put into place to handle the capacity of conducting classroom observations and utilizing findings. The strengths of the measure include high levels of validity, reliability, credibility, and meaningfulness. Without the proper resources devoted to the classroom observation process, one or more of the strengths completely disappear. Each formal observation cycle easily takes 12 or more hours, for the observer. Typically, there are two formal observation cycles in a year and several informal observations per teacher. Given the time required is considerable, the responsibility cannot be simply absorbed into other jobs. I recommend hiring a minimum of one full-time classroom observer for each school site to facilitate data collection and follow-up with teachers (actual number would depend on student enrollment and staffing). I think this is one area that needs to be more fully explored (mentioned below). In my opinion, an observer cannot implement classroom observations, with fidelity, for more than 60 teachers per year (the number may be closer to 50).

Not only is the time commitment considerable for observers, but also for teachers. Teachers easily spend 8 or more hours involved in each formal observation cycle. Given the formative role of observation, teachers also have additional considerable work (including professional development activities) following an observation cycle. In my experience working with educators, teachers need ample time to build knowledge before implementing changes in classroom practices. Changes in attitude and perspective, including knowledge, precede changes in behavior (and even more distal, student outcomes). It is not unreasonable to expect changes in classroom practice and student outcomes one or two years after the professional development effort (Guskey, 2006; Puma and Raphael, 2001). In order for the formative role of classroom observations to become fully realized, the cycle of educator learning and development must be respected. I recommend using classroom observations every other year for the assessment of teacher effectiveness. In my opinion, the focus should alternate between assessment and development. Teachers should be given adequate time to engage in professional learning opportunities, internalize the learning, and incorporate changes into their classrooms before being observed again. As this study shows, changes in teacher effectiveness take time. Too much observation will exhaust resources, cause evaluation burnout, and most importantly stifle the improvement process.

Lastly, concerning the use of observations, I believe that all components of an observation protocol should be incorporated in each formal observation cycle. In LAUSD, the district is identifying "focus elements" for a specified school year, which total approximately one third of the observable elements in the LAUSD Teaching and Learning Framework (LAUSD Talent Management Division, 2013). Limiting the focus in an observation to selected elements threatens the validity of the instrument and poses significant challenges for implementation. As this study finds, designing instruments is extremely difficult. Even an instrument with high face validity can be problematic. The process of ignoring elements in the observation framework alters the integrity of the instrument. While focusing on select elements may somewhat reduce the time it takes to code observation data, it also has the potential to introduce multiple negative unintended consequences into the system. Additionally, if focus elements shift on a yearly basis, the summative aspect of the evaluation is severely crippled as follow-up observations are centered on different elements. Considering that not all teachers are observed during a given year, the comparability of estimates of effectiveness across teachers and years is also limited. Valid and reliable observation protocols take years to develop and should be implemented in their entirety. I recognize that classroom observations take a tremendous amount of resources to implement with fidelity and I think that the effort is well worth it. In fully actualizing the promise of a CTES, a significant amount of resources are required, the bulk of which can be expected to go to classroom observations.

Implications for Improving Teacher Quality Professional Development Programs

This studies investigation into the multiple ways to measure teacher effectiveness not only has implications for comprehensive teacher evaluation systems but also for professional development programs geared at improving teacher quality. Through close examination of the teacher survey, expert assessment, and classroom observation data, important considerations for PD programs and the evaluation of those programs are revealed. In this section, I briefly address the most salient considerations that have been illuminated in this study.

Considerations for Professional Development Programs

There is a plethora of literature suggesting that diverse use of instructional strategies is a component of effective teaching (Darling-Hammond, et al., 2009; Elmore, 2002; Goe, et al., 2008; Kemp & Hall, 1992; National Board for Professional Teaching Standards, 1989). The context for this study – the ITQ Program, was structured around the belief that effective teachers use many available resources and call on multiple methods to teach their students and meet their goals (Goe, et al., 2008; National Board for Professional Teaching Standards, 1989). This includes, among other things, having a repertoire of teaching strategies (Darling-Hammond, 2011; Kemp & Hall, 1992). Findings from this study support the use of literacy instructional strategies as a way to support the needs of a diverse student population.

Both the expert assessment and classroom observations used in this study measured the use of broad sets of instructional strategies in the classroom. The expert assessment captured differences in use of instructional strategies between participation groups (higher APD participation teachers used more strategies) and changes in use of instructional strategies over time (increasing for moderate and high participation groups). The various teacher effectiveness constructs on the teacher survey are highly and often significantly correlated with frequency of instructional strategy use, as was APD participation. Similarly, the count instructional strategies used within a class period is highly and significantly correlated with scores on classroom environment and instruction. In fact, the number of strategies used within a class period explains 56% to 75% of the variation in scores on classroom observation teacher effectiveness constructs (based on the order in which predictors were entered into the final reported models). Findings from this study strongly suggest that the focus on building knowledge and use of instructional strategies is a worthwhile effort for professional development programs geared at improving teacher quality.

Considerations for the Evaluation of Professional Development Programs

Having been involved in both the evaluation of ITQP and this close examination of the measures used in ITQP, I have gained unique insights into the evaluation of improving teacher quality programs. One of the key lessons I have learned has to do with the design and use of evaluation instruments. All too often evaluators are put into the position of designing or tailoring an instrument to the evaluation needs of a particular program. Unfortunately, evaluators are also (more times than not), subject to strict program and funding timelines that do not allow for time to properly develop evaluation instruments. This can be disastrous for the evaluation of said programs (resulting in issues with validity, reliability, and difficulties concerning interpretation of findings).

In the case of ITQP, the funding of the evaluation began two months after the start of the program leaving no opportunity to collect baseline data and challenges in producing usable evaluation instruments well aligned with program goals. Fortunately, due to other work with the program provider, the Teacher Survey, used in this study, had already been in the process of development for several years (informed by multiple survey sources, informed by multiple stakeholder groups, reviewed by program evaluators at the international American Evaluation Association (AEA) annual conference, pilot tested, undergone exploratory factor analysis, etc.). Having spent a great deal of time thoughtfully and systematically developing and refining the instrument, it was deemed appropriate for use in the ITQP evaluation. Even though the teacher survey seemed to be a valid and reliable way to measure teacher effectiveness, in relation to engagement in high quality professional development, this study's findings suggest this may not be the case. This study proves just how difficult (and perhaps unwise) it is to design one's own instruments for use in program evaluation.

Creating evaluation instruments is both an art and science. There are plenty of guides highlighting key practices to constructing a good survey instrument (Bradburn, et al., 2004; Fowler, 1995; Sudman, et al., 1982; Trochim, 2006). Typical suggestions include how to organize the instrument (begin with an introduction, put items in logical order, place sensitive demographic items near the end, etc.), how to construct questions (use simple words, use concise sentences, choosing response scales, etc.), questionnaire formats (mode of administration, consideration for the respondent, etc.), and what not to do (don't use double negatives, don't write leading questions, etc.) (Trochim, 2006). For those serious about constructing good instruments, there are also plenty of suggestions on how to test and refine the validity and reliability of the instrument, including field testing (piloting surveys), proper training of interviewers and observers, eliciting respondent feedback (cognitive interviews, behavior coding, response latency, formal debriefings, vignette analysis), debriefing of interviewers, triangulation with other data sources, and use of various statistical procedures (latent class analysis (LCA), item response theory (IRT), multitrait multimethod matrix (MTMM), and other complex modeling procedures) (Presser, et al., 2004; Trochim, 2006; Willis, 1999).

Designing and refining good instruments takes a lot of time and resources. Even when done well, there are many potential sources of measurement error (both random and systematic) in survey methodology (Alwin, 2007; Biemer et al., 2003; Biemer et al., 2013; Viswanathan, 2005). Being able to identify the presence and extent of measurement error improves both data collection and analysis (Alwin, 2007; Biemer, et al., 2013). Ultimately, this study shows is that even when adhering to guidelines on developing and refining instruments, fatal flaws may persist. My recommendation for program evaluators of improving teacher quality programs is to use preexisting validated instruments whenever possible. The time involved in locating instruments is far less than the time it will take later to attempt to correct for measurement error.

Study Limitations

There are multiple limitations to consider when interpreting the results and implications of this study. Limitations span from study design, to instrumentation and analysis, to school and district context, to study sample. In many cases, limitations are inherited through use of ITQP evaluation data. The relevant set of concerns is addressed under each topic area.

Study Design

One limitation of this study concerns the total number of measures involved in the study. Given the host of potential measures for inclusion in a CTES (e.g. self-reported survey data, expert assessment, classroom observation, video assessment, parent feedback, student feedback, student work, peer-review, self-evaluation, student achievement data, etc.), deep exploration of three measures, while important, only begins to contribute the necessary knowledge base on measuring teacher effectiveness for use in a CTES. While higher than most other published studies (often relying on one measure of teacher effectiveness), there still remain multiple measures for study in CTESs.

The most complete study would involve longitudinal data collection on all possible measures of teacher effectiveness on a sizable group of teachers. When all the measures considered for inclusion in a CTES are studied simultaneously, a complete understanding of what each measure of teacher effectiveness contributes to an overall understanding of teacher quality (and how each measure relates to another) would be discernable. Although having the full range of measures contained in one study is ideal, it is also not feasible, given time and resource constraints.

A deeper limitation for this study is that not all three measures of teacher effectiveness are available for the full extent of the study. While longitudinal data is available for the teacher survey data (3 years) and to a lesser extent for the expert assessment data (2 years), classroom observation data is only available for 1 year. This limits the true ability to compare depictions of teacher effectiveness across measures (as it is only fully available in year 3). The restricted time frame for expert assessment data and classroom observation data also hinders the ability to fully answer research question one regarding sensitivity to detecting between-group and within-group differences. The lack of baseline data (pre ITQP) for any measure makes answering questions about actual change over time, differences between participation groups, and impact of academic PD participation through ITQP challenging.

Lastly, there were host of other competing interventions occurring simultaneously as ITQP at the school sites. It is unknown to what extent teachers participated in those activities, all of which could have had an impact on teacher effectiveness. Given the impossibility of tracking teacher participation in outside ITQP activities, the formation of "groups" strictly based on ITQP APD may or may not be the best method to explore each measure's sensitivity level to detecting differences. There may be in fact other more defining factors that differentiate teachers from one another, currently unknown. Simple maturation effect also plays an unknown role in the study.

Instrumentation and Analysis

The data sources used in this study each pose a unique set of limitations. Although proven to be a valid and reliable instrument (through exploratory and confirmatory factor analysis), the teacher survey is subject to the inherent biases present in self-reported data. Teachers indicate their level of agreement with attitudinal items and the frequency with which they engage with various behavioral indicators, related to teacher effectiveness. Teacher responses on attitudinal items are higher than on behavioral items suggesting some social desirability in responses. Because of the self-reported nature of the teacher survey, it is not an objective measure of teacher effectiveness and therefore limits inferences that can be made about effectiveness and overall teacher quality. Findings surrounding significant declines in ongoing learning for all three participation groups reveal threats to the internal validity of the instrument. The overall decline in ongoing learning is thought to be a direct reflection on district budget cuts that made it extremely difficult for teachers to take advantage of conference activities, the primary source of interest in the ongoing learning construct. The overall decline in ongoing learning is less of a measure of teacher effectiveness (internal ability) than it is a temperature reading for the district enabling teachers to take advantage of conference and other PD opportunities (external factors). The within-group differences identified for the ongoing learning construct are not convincing for this studies purpose, leaving the increases in reflective practice as the only meaningful change in teacher effectiveness (according to the teacher survey) over the three years. Pattern analysis findings from this study also suggest that the teacher survey is overly prone to influence by outside variables, limiting its use in a CTES.

The expert assessment was only used in year 2 and 3 of the study and therefore limits the ability to test for significant change over time in construct scores. Furthermore, the expert assessment only captures frequency of use of instructional strategies. Although frequency of use is important, quality of use is arguably of greater relevance. Due to the non-evaluative nature of the PLP relationship with teachers, the focus of the expert assessment remains centered on frequency of use, which is thought to be a precursor to quality. At some point however, one would expect to see an effective teacher using fewer strategies because s/he is utilizing a strategy to its fullest. An additional concern with the expert assessment is the introduction of a conflict of interest. PLPs lead ITQP PD focused on increasing knowledge and use of instructional strategies. Given PLPs have a vested interest in seeing teachers use more instructional strategies they may not be the most objective sources to assess actual instructional strategy use, which may be why there is a lack of alignment between with findings from classroom observations.

The teacher observation protocol, available for the third year of the study, affords little insight into change over time or differences amongst participation groups. There is some evidence to suggest that for the high participation group, there may be some growth in construct scores. Ultimately, scores for teachers with two observations are averaged, in an attempt to study classroom practice in the most rigorous way. In this case, a baseline reading would have been very enlightening. Moreover, due to the nature of negotiating access to classrooms, the majority of teachers participating in classroom observations are in the moderate and high participation groups further limiting the ability to test the measure's sensitivity to detecting between-group differences. The ability to make comparisons across measures is severely limited due to the single year administration.

In the cases where multiple tests of significance are conducted, the necessary statistical adjustments are performed to counteract the error. In my opinion, the adjustments made are adequate for this study. Most likely, this is not a problem because where there is statistical significance; the levels are very high which suggests any additional adjustments would not make substantial differences.

School and District Context

In general, teachers at school's A and B were faced with a plethora of challenges ranging from low operating budgets, high administration turnover, high teacher turnover, district teacher evaluation changes, campus crime, lawsuits, School Improvement Grant (SIG) requirements, etc. The multitude of challenges faced by teachers likely impacted their level of effectiveness and undoubtedly impacted their ITQP APD participation levels. School and district challenges also created a hesitancy to participate in data collection efforts stemming from low morale, overwhelm by other research activities, and fear surrounding use of findings. Each of these challenges has an impact on the study, from the ability to form sizable participation groups, to accessing data, to making comparisons across measures.

Study Sample

Overall, the teachers in this study score lowly across the objective measures of teacher effectiveness. School's A and B are known nationally for their perpetual low-performance, which is a primary reason the schools were targeted for improvement by the ITQ Program. The high-need classification of both the student and teacher populations is directly related to the individual and collective estimates of teacher effectiveness and overall teacher quality. In practical terms, this means that all groups are scoring low across measures of teacher effectiveness, which makes it hard to detect changes between groups and/or over time. On the attitudinal items on the teacher survey, teachers score very high and remain high throughout the study, which also makes it hard to detect changes between groups and/or over time. The populations of teachers and students included in the study lack the diversity required for adequate representativeness, thereby limiting the studies generalizability.

Directions for Future Research

There are many relevant directions research on building comprehensive teacher evaluation systems can take. This study aids in developing an understanding of several potential measures for inclusion in a CTES, namely a teacher survey, expert assessment, and observation protocol. At the commencement of this study there was still so much to be learned about selecting measures for inclusion in a CTES. In the past few years some progress has been made on this front. The findings from this study add to that knowledge base and help narrow the field of potential measures for inclusion in a CTES. One possibility for continuing research is to replicate a longitudinal study with alternative measures of teacher effectiveness to get a more in-depth view of what each additional measure of teacher effectiveness can contribute to a complete understanding of teacher quality. Similar research questions can be addressed creating an even deeper insight into how best to design a CTES.

A recent study by the Measures of Effective Teaching (MET) Project, funded by the Bill & Melinda Gates Foundation, set out to build and test measures of effective teaching, which could then be used by districts to identify and develop great teaching (MET website, 2014). A three-year study (which took place concurrently with data collection for this study: 2009-2012) identified three distinct measures of teacher effectiveness (classroom observations, student perception surveys, and measures of student achievement) that collectively assess overall teacher effectiveness in a fair and reliable way (MET Project, 2013). The study does an excellent job exploring each measure's contribution to assessing effective teaching. As a more complete study, I think the three measures identified by the MET Project and the lessons learned should set the foundation for continuing research on developing CTESs.

What this study and the study conducted by the MET Project have in common is that they identify classroom observations as an integral component of a CTES. In order to conduct classroom observations as proposed by experts in the field (multiple observations (both formal and informal), well documented (lesson transcripts, coding of evidence to support ratings, notifications to teachers), pre-post observation conferences, review of lesson plans, timely turn-around, etc.) (Danielson, 2007; LAUSD, 2013; MET Project, 2013; UTLA, 2013), new ideas about how to logistically conduct teacher evaluation via observation are needed.

In traditional teacher evaluation models involving personnel evaluation, an administrator (usually the principal), is responsible for conducting the evaluation. Whereas traditional observations are typically short and infrequent, with little documentation, the new observation models are quite complex and time consuming (as mentioned above). There is no way that the burden of conducting personnel evaluation should rest on the shoulders of administrators alone. I do think administrators should be involved to an extent in the process but in reality a new way of thinking about collecting observation data needs to be paired with the new observation protocols and cycles. This likely means creating many new positions for certified observers, trained in the observation cycle and privy to professional development opportunities for developing teachers.

An area that needs to be more fully explored is who conducts the observations and how a particular observer impacts the formative and summative aspects of the evaluation. Does the observer need to be an accomplished teacher? A content expert? An instructional coach? A researcher? A professional development expert? What background characteristics are ideal and/or minimal for valid and credible ratings? What sort of training does the observer require (to develop inner-rater reliability) and what sort of follow-up (observer calibration) is needed to maintain accuracy and reliability? What support do observers need in connecting sub-optimal evaluation ratings with proven successful targeted professional development? How many teachers can an observer evaluate and support in a given year? How many observers are needed for a school/district/etc.? These are all questions that are of immediate importance to answer as new CTESs are implemented. If we do not create the infrastructure to implement CTESs with fidelity, failure is eminent.

Continuing with this line of reasoning, further research is needed surrounding the knowledge base of effective professional development programs. In order for a CTES to fulfill its purpose of improving teaching and learning, knowledge and systems must be in place to efficiently connect effective professional development with a teacher's identified area(s) in need of development. At the moment, there is no comprehensive menu of professional development programs nor do we have a firm grasp on what particular PD is actually successful at improving various aspects of effective teaching. A CTES needs structures in place to handle redirection of teachers once evaluation ratings are complete. This requires development of the knowledge base surrounding professional development programs, administrative capacity to connect teachers to optimal professional development, and the resources to support teachers in engaging in the targeted professional development. Of immediate importance is systematically identifying effective professional development programs, with full knowledge of the expected impact on teaching for each PD opportunity.

Proponents of CTES firmly believe that a measure of student achievement is an integral component in a multi-measure system. While the need to include a measure of student achievement is widely recognized, there is still much discussion around what this looks like. For many subject areas (e.g. social studies-history, languages, art, music) standardized testing is irregular or completely nonexistent. Much more research is needed on alternative student achievement measures and the applicability of those measures across the elementary, middle, and high school continuum.

Lastly, there is the pressing issue of the proper weighting of measures in a CTES. The MET Project tested four different weighting models and recommends either assigning 50% (with 25% assigned to each remaining measure) or 33% (equal weighting of 3 measures) of the weight to a measure of student achievement (MET Project, 2013). While their study preliminarily establishes the tradeoffs of various weighting models, much still remains to be learned as states move forward with various weighting schemas. Ultimately, a composite teacher quality score is to be discerned from the multiple-measures of teacher effectiveness in a CTES. Weights assigned to each measure directly impact a teacher quality composite score, ultimately influencing the summative function of the evaluation. In the next few years it is absolutely imperative that studies are undertaken that closely look at the depictions of teacher quality based on CTES measures and weights and the overall effect on the educational system.

Final Remarks

Knowing what we value in education should guide what we seek to measure. Once we know the goals that we seek to achieve, it is our moral imperative that we measure the achievement of those goals in accurate, valid, reliable, credible, and fair ways. If one of our goals is to have an effective teacher in every classroom in America, then we need to build systems that help us get there. Well-designed comprehensive teacher evaluation systems (CTESs) can help identify, retain, reward, and develop effective teachers, ultimately reinvigorating public education.

If we are to move forward with implementing CTESs, we must be prepared to use findings appropriately. This means that society must be in the position to act on findings in such a way that honors the two primary goals of such a system: to improve teaching and learning and to make sound personnel decisions. In order for this to happen a lot has to change. Unilaterally, schools, districts, states, and the federal government must position themselves to both implement CTESs (with fidelity) and use findings (appropriately) from such a system. An educator growth and development cycle, its crux being evaluation, must be in place.

Formatively using teacher evaluation findings first entails knowing interventions that are known to affect change in weak areas targeted for improvement. Beyond knowledge of those interventions, there has to be funding to support teachers in choosing and pursing professional development related to those areas. Ideally, there would be a menu of successful professional development that teachers can select from to personalize their personal professional growth and development. Teachers would be given adequate time to develop in selected areas and then reassessed for growth at a reasonable time in the future.

Using teacher evaluation findings summatively entails first having the power to act on knowledge. There are too many roadblocks to accessing quality education, from state laws to teacher unions that prevent sound decision-making. The Vegara v. California case is a perfect example of such entrenched challenges. In order for progress to materialize, administrators must be in the position to staff their schools with the number one consideration in mind being what is best for their students. Teachers need to demonstrate successful performance (and be given adequate time to get there) before being guaranteed a job for life. Perpetually ineffective teachers must be let go (without a fight that costs hundreds of thousands of dollars and years in litigation). Ultimately, trust has to be infused into a system where all parties are united in the goal of ensuring access to quality education.

With such a system, teacher evaluation becomes truly meaningful both for the teacher and for society as a whole. When we can guarantee that effective teachers teach our children, we can expect not only high academic achievement but also a lifetime of success. As the federal government, states, and districts, are proposing new teacher evaluation systems with inherently higher levels of validity and reliability that can more accurately and meaningfully assess teacher effectiveness, all stakeholders must find a way to support the transition to new CTESs. Discussions need to revolve around the "right" measures for inclusion and how we proceed with using such measures and systems.

Through the use of multiple-measures to assess effective teaching, comprehensive teacher evaluation systems take a holistic approach to teacher evaluation recognizing and valuing the complexities of teaching, differentiating levels of teaching, and advocating for effective teaching. By honoring both teaching and learning, the teaching profession is held in high regard and every child's constitutional right to quality education is protected. I firmly support the belief that teaching matters and hence would like to see successful systems in place that protect the sacredness of both teaching and learning. Findings from this study help those involved in the many capacities of teacher evaluation, especially those involved in thoughtfully building and progressing comprehensive teacher evaluation systems.

Appendix A
Description of the Improving Teacher Quality Program (ITQP) and Evaluation

The Partnership for Los Angeles Schools (PLAS) and the University of California Los Angeles (UCLA) Center X worked together to provide Professional Development (PD) and support to science and social studies teachers at two middle schools (serving 6th, 7th, and 8th grade students) during the 2009-2010, 2010-2011, and 2011-2012 academic years. The work focused on closing the Achievement Gap for African-American and Latino students, English Language Learners, and Students with Disabilities in science and social studies-history. The California Postsecondary Education Commission (CPEC), now a part of the California Department of Education (CDE) provided funding for this project through an Improving Teacher Quality (ITQ) Grant. The ITQ Program logic model is depicted in Figure A-1. The specific goals of the program included:

Short Term:

1. Support teachers in their development of content knowledge and their knowledge/use of instructional strategies for a diverse student population.

Medium Term:

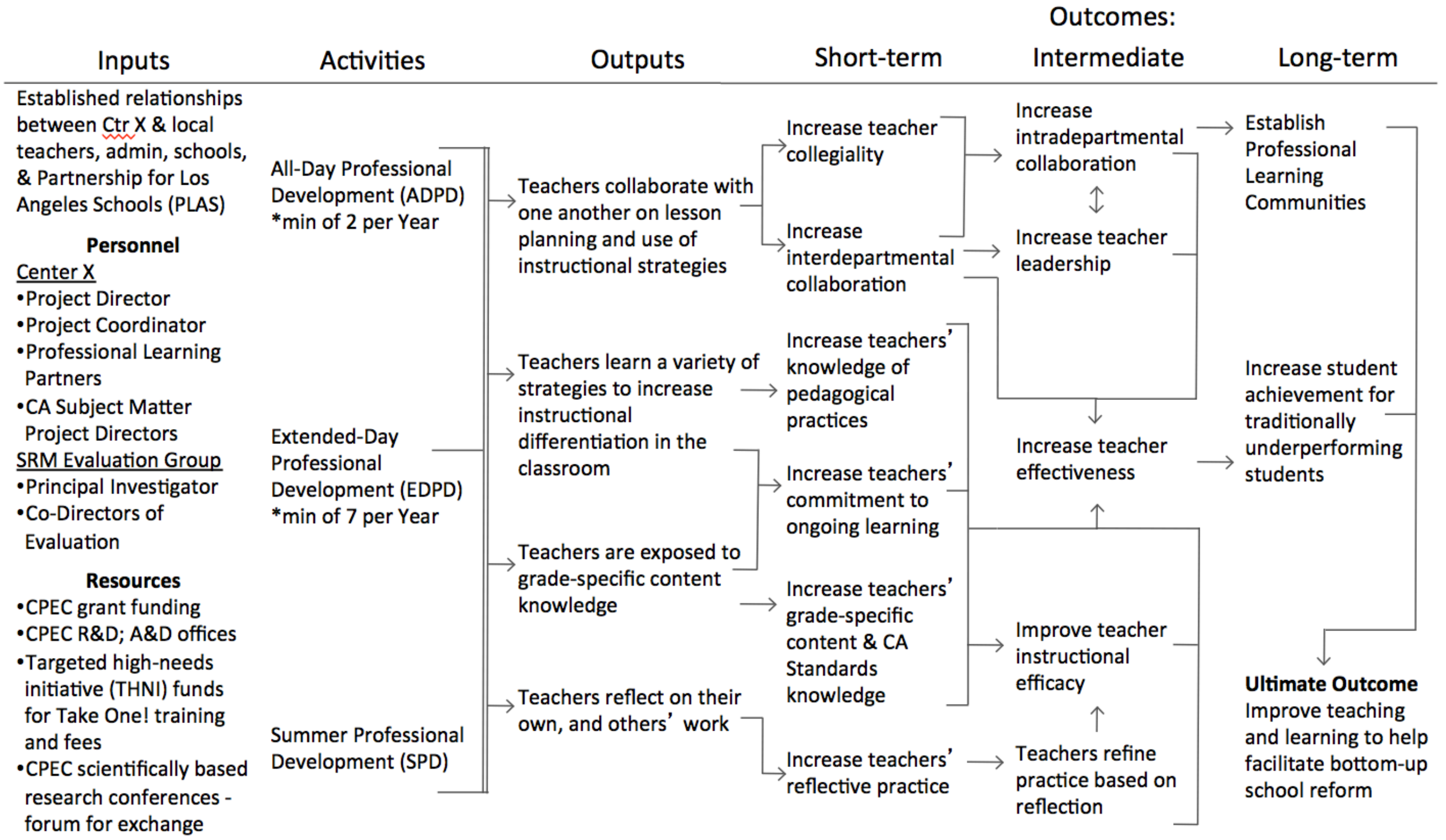
2. Increase teacher effectiveness.

Long Term:

3. Establish Professional Learning Communities (PLCs).
4. Increase student achievement in science and social studies.

Figure A-1.

Improving Teacher Quality Program Logic Model



Outcomes:

Inputs | Activities | Outputs | Short-term | Intermediate | Long-term

Inputs

Established relationships between Ctr X & local teachers, admin, schools, & Partnership for Los Angeles Schools (PLAS)

Personnel

Center X

- Project Director
- Project Coordinator
- Professional Learning Partners
- CA Subject Matter Project Directors

SRM Evaluation Group

- Principal Investigator
- Co-Directors of Evaluation

Resources

- CPEC grant funding
- CPEC R&D; A&D offices
- Targeted high-needs initiative (THNI) funds for Take One! training and fees
- CPEC scientifically based research conferences - forum for exchange

Activities

All-Day Professional Development (ADPD) *min of 2 per Year

Extended-Day Professional Development (EDPD) *min of 7 per Year

Summer Professional Development (SPD)

Outputs

Teachers collaborate with one another on lesson planning and use of instructional strategies

Teachers learn a variety of strategies to increase instructional differentiation in the classroom

Teachers are exposed to grade-specific content knowledge

Teachers reflect on their own, and others' work

Short-term

Increase teacher collegiality

Increase interdepartmental collaboration

Increase teachers' knowledge of pedagogical practices

Increase teachers' commitment to ongoing learning

Increase teachers' grade-specific content & CA Standards knowledge

Increase teachers' reflective practice

Intermediate

Increase intradepartmental collaboration

Increase teacher leadership

Increase teacher effectiveness

Improve teacher instructional efficacy

Teachers refine practice based on reflection

Long-term

Establish Professional Learning Communities

Increase student achievement for traditionally underperforming students

Ultimate Outcome

Improve teaching and learning to help facilitate bottom-up school reform

Assumptions:

- Increases in teacher effectiveness lead to better teaching and improvement in student performance on standardized achievement measures.
- The development of small, collaborative teacher groups will lead to the establishment of Professional Learning Communities.
- Programmatic effects will translate beyond the departments to the whole school.

External Factors:

The following may impact program success:

- School characteristics & demographics (i.e. admin, school culture, district & PLAS context)
- Teacher turnover may limit potential program effects
- District context (including political and legal issues)
- School-site funding for coaches

Ultimate Goal:

5. Improve teaching and learning to help facilitate bottom-up school reform.

Improving Teacher Quality Program (ITQP) Design

The ITQ Program used an *Understanding by Design* (Wiggins and McTighe) model to help identify, plan, and implement curriculum that supports learning for diverse student populations. ITQP also used the *Response to Intervention (RTI)* model to assess student learning to drive instruction and professional development. Furthermore, *Adaptive Schools* training helped foster leadership and collaboration at school sites. The *Cognitive Coaching* model informed delivery of all program services and was used to support teachers in their growth as professionals, regularly engaging in reflective practice.

The program consisted of UCLA's Center X Professional Learning Partners (PLPs) providing intensive professional development to all teachers in science and social studies at two LAUSD middle schools (School A and School B). The professional development plan for both sites offered many different learning opportunities that supported teachers in their development and understanding of content, pedagogy, and leadership. The program was also designed in such a way as to foster the creation of sustainable Professional Learning Communities.

Specific components of the professional development included:

- **All Day PD (ADPD)** – This PD focused on developing content knowledge, pedagogy, and curriculum for a diverse student population, including English Language Learners (ELLs) and Students with Disabilities (SWDs). Using a Culturally Relevant and Responsive Education (CRRE) and following the Understanding by Design (UbD) approach, UCLA History/Geography Project, UCLA

Science Project, and UCLA Center X Professional Learning Partners (PLPs or instructional coaches) led the PD. ADPD lasted 7 hours and was generally held on weekends.

- **Extended Day PD (EDPD)** – PLPs led these PD sessions, which focused on building knowledge and use of instructional strategies, collaboration, and the application of learning from ADPD. EDPD took place after school-wide departmental common planning time and lasted 1.5 hours.
- **Capacity Building Summer Institutes** – These weeklong institutes occurred after the first and second academic years of the program. UCLA History/Geography Project, UCLA Science Project, and UCLA Center X PLPs led the institutes, which focused on teaching and developing inquiry-based lessons, deepening content knowledge, leadership building, facilitation skills, developing cultures of collaboration, and planning for future professional development.
- **Peer Coaching and Classroom Observations** – Training and guidance was provided by on-site PLPs during "friendly visits" that included non-evaluative data collection, observations, and analysis of student work. These focused on developing pedagogical practices, and building PLCs and collaboration.
- **Take One! Lesson Study (Year 3)** – Teachers who participated in the program in the third year videotaped lessons as a part of the National Board Certification Process. The APD surrounding this focused on incorporating instructional strategies, Understanding by Design, Response to Instruction and Intervention, and community building into lesson design.

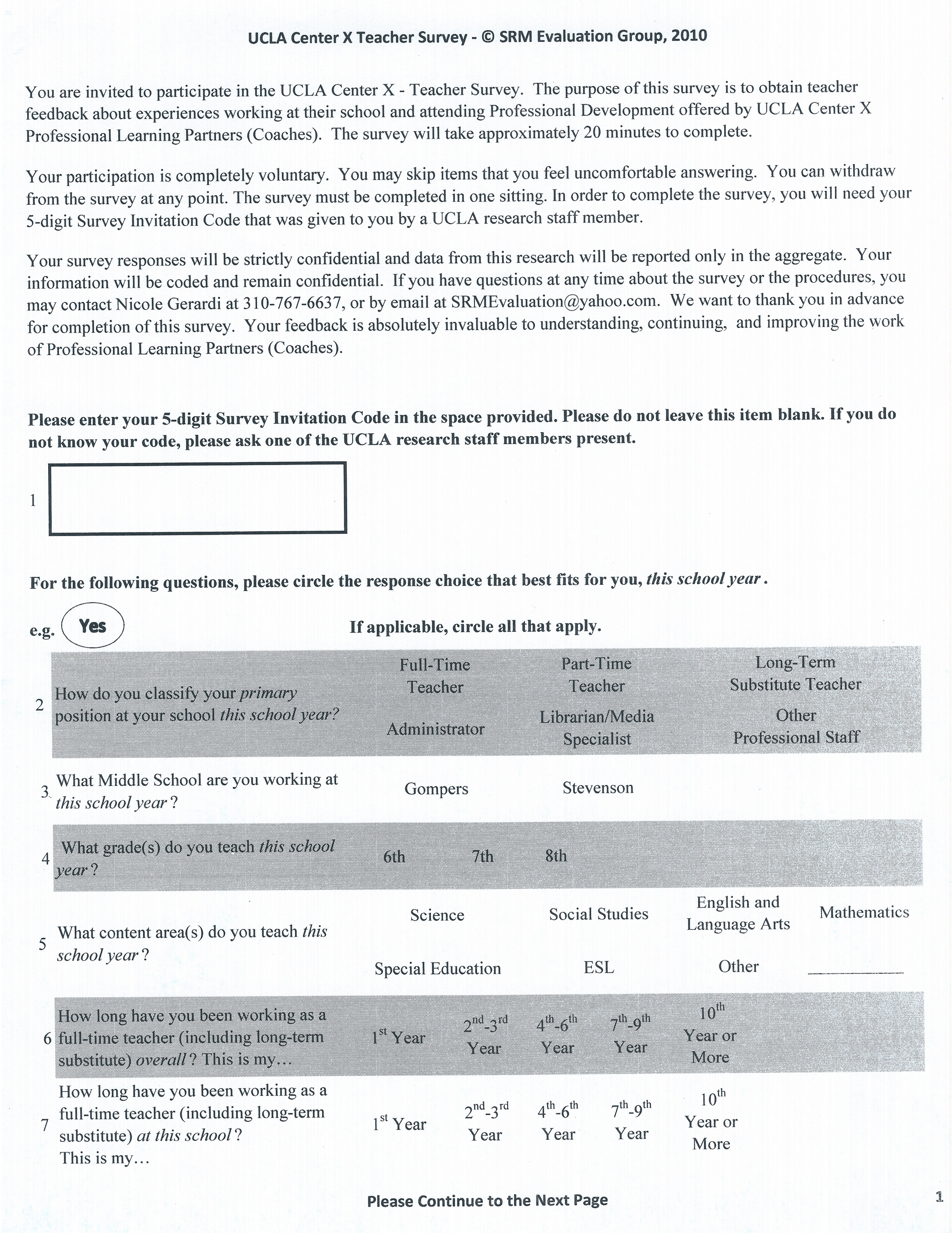
Improving Teacher Quality Evaluation

The evaluation of the Improving Teacher Quality Program involved multiple measures used within a quasi-experimental design to assess the degree to which the program met its primary goals for teachers and students. The evaluation consisted of both formative and summative assessment, and measured both program implementation and outcomes. Specifically, the following primary research questions were addressed:

1. In what ways does participation in ITQP impact Teacher Effectiveness?
2. In what ways does participation in ITQP contribute to the establishment of Professional Learning Communities?
3. In what ways do participating teachers differ from non-participating teachers in their classroom practice?
4. Do student achievement scores increase at a greater rate in treatment schools versus comparison schools?

For a more in depth view on the program, evaluation, and findings, please see the final evaluation report by MacCalla, Dillman, & Alkin (2013), prepared for the California Department of Education (CDE).

Appendix B
Teacher Survey©



UCLA Center X Teacher Survey - © SRM Evaluation Group, 2010

You are invited to participate in the UCLA Center X - Teacher Survey. The purpose of this survey is to obtain teacher feedback about experiences working at their school and attending Professional Development offered by UCLA Center X Professional Learning Partners (Coaches). The survey will take approximately 20 minutes to complete.

Your participation is completely voluntary. You may skip items that you feel uncomfortable answering. You can withdraw from the survey at any point. The survey must be completed in one sitting. In order to complete the survey, you will need your 5-digit Survey Invitation Code that was given to you by a UCLA research staff member.

Your survey responses will be strictly confidential and data from this research will be reported only in the aggregate. Your information will be coded and remain confidential. If you have questions at any time about the survey or the procedures, you may contact Nicole Gerardi at 310-767-6637, or by email at SRMEvaluation@yahoo.com. We want to thank you in advance for completion of this survey. Your feedback is absolutely invaluable to understanding, continuing, and improving the work of Professional Learning Partners (Coaches).

Please enter your 5-digit Survey Invitation Code in the space provided. Please do not leave this item blank. If you do not know your code, please ask one of the UCLA research staff members present.

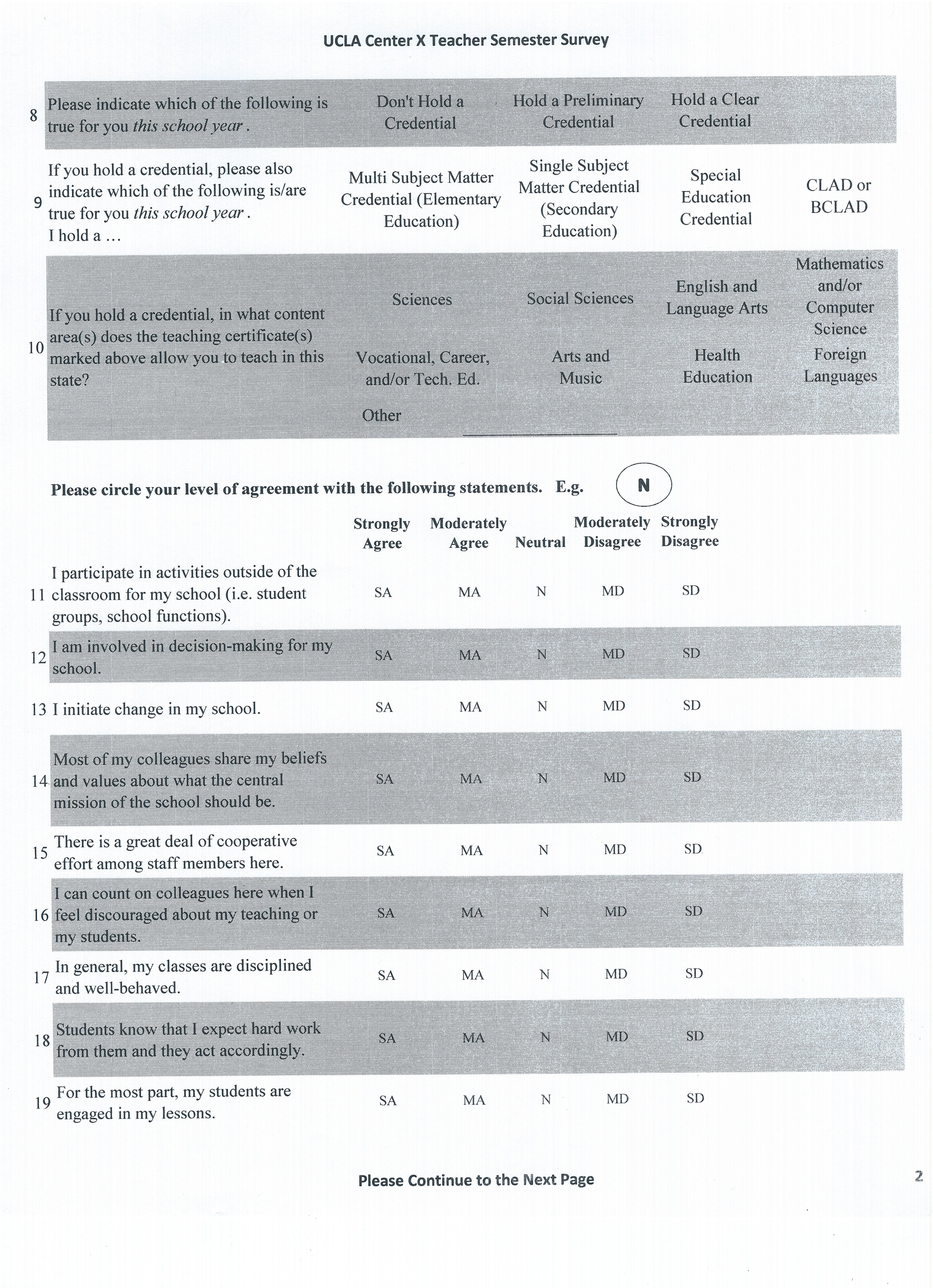
1 □

For the following questions, please circle the response choice that best fits for you, *this school year*.

e.g. [✓ **Yes**] **If applicable, circle all that apply.**

2	How do you classify your *primary* position at your school *this school year?*	Full-Time Teacher Administrator		Part-Time Teacher Librarian/Media Specialist	Long-Term Substitute Teacher Other Professional Staff	
3	What Middle School are you working at *this school year*?	Gompers		Stevenson		
4	What grade(s) do you teach *this school year*?	6th	7th	8th		
5	What content area(s) do you teach *this school year*?	Science Special Education		Social Studies ESL	English and Language Arts Other	Mathematics __________
6	How long have you been working as a full-time teacher (including long-term substitute) *overall*? This is my…	1st Year	2nd-3rd Year	4th-6th Year	7th-9th Year	10th Year or More
7	How long have you been working as a full-time teacher (including long-term substitute) *at this school*? This is my…	1st Year	2nd-3rd Year	4th-6th Year	7th-9th Year	10th Year or More

Please Continue to the Next Page



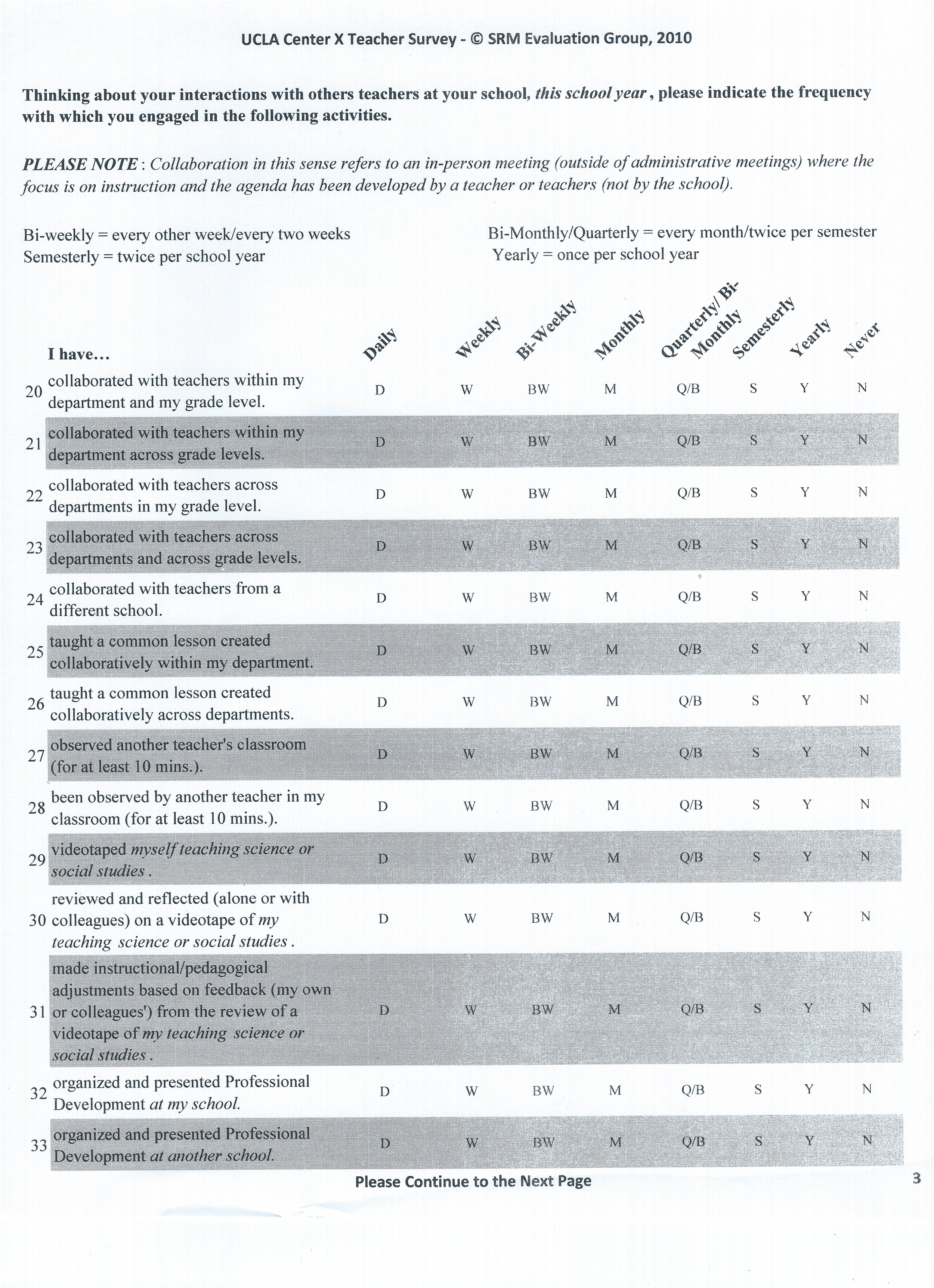
UCLA Center X Teacher Semester Survey

8	Please indicate which of the following is true for you *this school year* .	Don't Hold a Credential	Hold a Preliminary Credential	Hold a Clear Credential	
9	If you hold a credential, please also indicate which of the following is/are true for you *this school year* . I hold a …	Multi Subject Matter Credential (Elementary Education)	Single Subject Matter Credential (Secondary Education)	Special Education Credential	CLAD or BCLAD
10	If you hold a credential, in what content area(s) does the teaching certificate(s) marked above allow you to teach in this state?	Sciences	Social Sciences	English and Language Arts	Mathematics and/or Computer Science
		Vocational, Career, and/or Tech. Ed.	Arts and Music	Health Education	Foreign Languages
		Other ______________			

Please circle your level of agreement with the following statements. E.g. N

		Strongly Agree	Moderately Agree	Neutral	Moderately Disagree	Strongly Disagree
11	I participate in activities outside of the classroom for my school (i.e. student groups, school functions).	SA	MA	N	MD	SD
12	I am involved in decision-making for my school.	SA	MA	N	MD	SD
13	I initiate change in my school.	SA	MA	N	MD	SD
14	Most of my colleagues share my beliefs and values about what the central mission of the school should be.	SA	MA	N	MD	SD
15	There is a great deal of cooperative effort among staff members here.	SA	MA	N	MD	SD
16	I can count on colleagues here when I feel discouraged about my teaching or my students.	SA	MA	N	MD	SD
17	In general, my classes are disciplined and well-behaved.	SA	MA	N	MD	SD
18	Students know that I expect hard work from them and they act accordingly.	SA	MA	N	MD	SD
19	For the most part, my students are engaged in my lessons.	SA	MA	N	MD	SD

Please Continue to the Next Page



UCLA Center X Teacher Survey - © SRM Evaluation Group, 2010

Thinking about your interactions with others teachers at your school, *this school year* , please indicate the frequency with which you engaged in the following activities.

PLEASE NOTE : *Collaboration in this sense refers to an in-person meeting (outside of administrative meetings) where the focus is on instruction and the agenda has been developed by a teacher or teachers (not by the school).*

Bi-weekly = every other week/every two weeks
Semesterly = twice per school year

Bi-Monthly/Quarterly = every month/twice per semester
Yearly = once per school year

	I have...	**Daily**	**Weekly**	**Bi-Weekly**	**Monthly**	**Quarterly/ Bi-Monthly**	**Semesterly**	**Yearly**	**Never**
20	collaborated with teachers within my department and my grade level.	D	W	BW	M	Q/B	S	Y	N
21	collaborated with teachers within my department across grade levels.	D	W	BW	M	Q/B	S	Y	N
22	collaborated with teachers across departments in my grade level.	D	W	BW	M	Q/B	S	Y	N
23	collaborated with teachers across departments and across grade levels.	D	W	BW	M	Q/B	S	Y	N
24	collaborated with teachers from a different school.	D	W	BW	M	Q/B	S	Y	N
25	taught a common lesson created collaboratively within my department.	D	W	BW	M	Q/B	S	Y	N
26	taught a common lesson created collaboratively across departments.	D	W	BW	M	Q/B	S	Y	N
27	observed another teacher's classroom (for at least 10 mins.).	D	W	BW	M	Q/B	S	Y	N
28	been observed by another teacher in my classroom (for at least 10 mins.).	D	W	BW	M	Q/B	S	Y	N
29	videotaped *myself teaching science or social studies* .	D	W	BW	M	Q/B	S	Y	N
30	reviewed and reflected (alone or with colleagues) on a videotape of *my teaching science or social studies* .	D	W	BW	M	Q/B	S	Y	N
31	made instructional/pedagogical adjustments based on feedback (my own or colleagues') from the review of a videotape of *my teaching science or social studies* .	D	W	BW	M	Q/B	S	Y	N
32	organized and presented Professional Development *at my school.*	D	W	BW	M	Q/B	S	Y	N
33	organized and presented Professional Development *at another school.*	D	W	BW	M	Q/B	S	Y	N

Please Continue to the Next Page

UCLA Center X Teacher Semester Survey

When thinking about the *past full year*, please indicate how many times the following has happened (this year), or will happen (next year). (Please include summer months)

PLEASE NOTE: *Professional Conferences require paid registration and travel to another site to attend or present a session at a local, state, or national meeting. Examples of professional conferences include: California Mathematics Council, California Association of Teachers of English, California Sciences Teachers Association, California Council for Social Studies, American Educationl Research Association (AERA), etc. Please refer to the number of unique conferences attended (e.g. one conference that lasts for three days should be counted as one conference).*

		0 times	**1 time**	**2 times**	**3 times**	**4 times**	**5 times**	**6 times or more**
34	In the *past 12 months*, I have *attended* a professional conference.	0	1	2	3	4	5	6+
35	In the *past 12 months*, I have been the *presenter* at a professional conference.	0	1	2	3	4	5	6+
36	In the *next 12 months*, I plan to *attend* a professional conference.	0	1	2	3	4	5	6+
37	In the *next 12 months,* I plan to *present* at a professional conference.	0	1	2	3	4	5	6+

From the list of topics below, please indicate the three that are your current top priorities for YOUR OWN Professional Development. ***In the line provided please handwrite your 1st, 2nd, and 3rd priority.***

38 Student Discipline and Classroom Management ____________________

39 Teaching Students with Special Needs ____________________

40 Teaching Students with Limited-English Proficiency ____________________

41 Use of Technology in Instruction ____________________

42 The Content of the Subject(s) I Primarily Teach ____________________

43 Content Standards in the Subject(s) I Primarily Teach ____________________

44 Methods of Teaching ____________________

45 Student Assessment ____________________

46 Communicating with Parents ____________________

Please Continue to the Next Page

For the following questions, please circle the response choice that best fits for you.

47	Are you certified by the National Board for Professional Teaching Standards in at least one content area?		No	Yes, Fully Certified	
48	Please indicate your gender.	Male	Female		
		African-American	American Indian/ Alaskan Native	Asian	Filipino
49	Please indicate your ethnicity.	Latino	Pacific Islander	White/ Non-Hispanic	
		Other ________________			

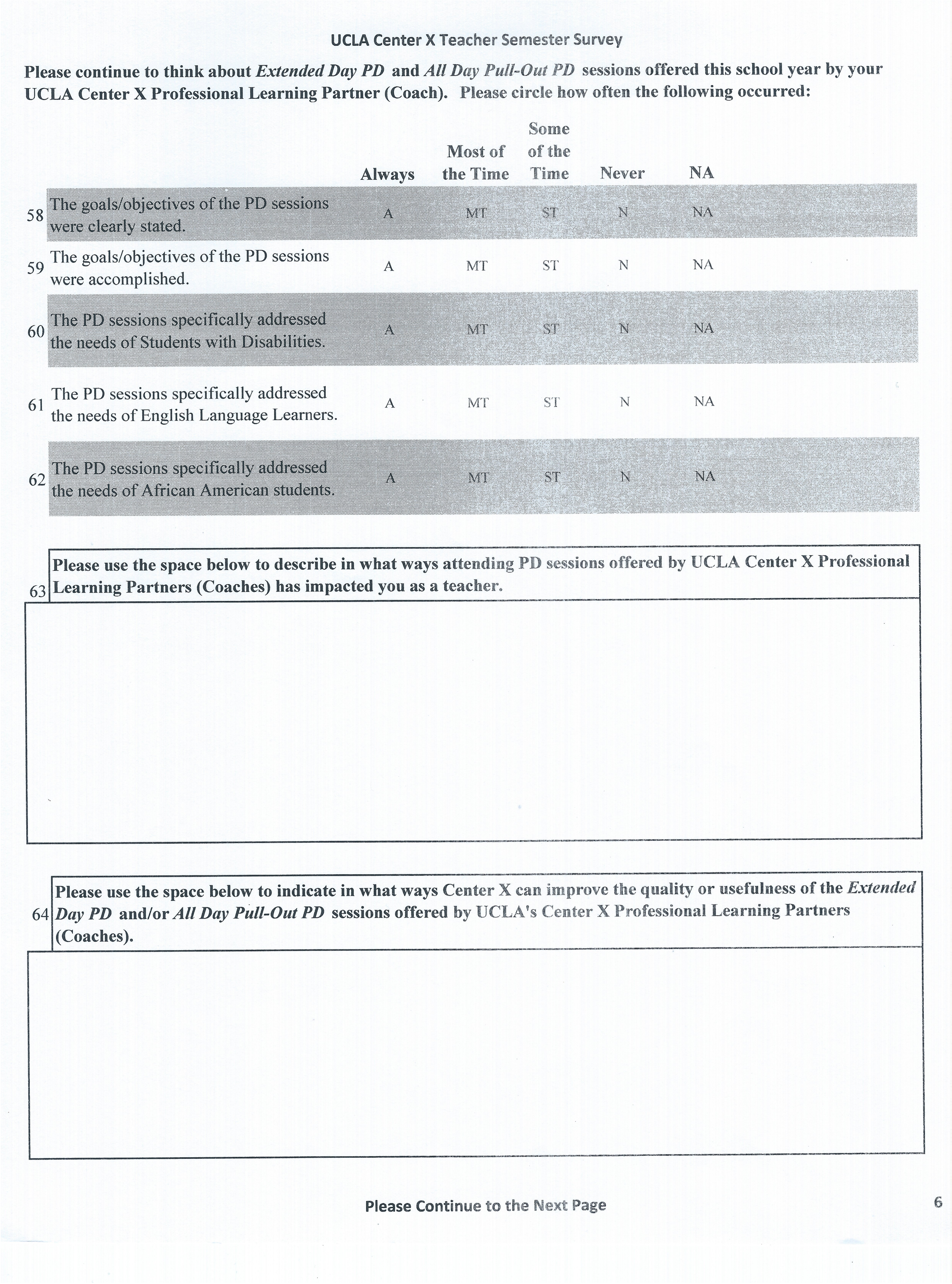
PLEASE NOTE: *Extended Day PD was held after school and usually lasts from 1.5-2 hours. The All Day Pull-Out PD was offered on Saturdays and was offered for both Gompers and Stevenson Middle Schools. It is also possible that the All Day Pull-Out PD was offered during the weekday and substitute teachers were provided for your classes. Both the Extended Day PD and the All Day Pull-Out PD were offered by the UCLA Center X Professional Learning Partners (Coaches).*

50	Have you attended any of the *Extended Day PD* or *All Day Pull-Out PD* sessions offered by a UCLA Center X Professional Learning Partner (Coach) *this school year*?	No	Yes	
If yes, …				
51	in what content area(s) have you been attending UCLA Center X *Extended Day* or *All Day Pull-Out PD this school year*?	Science	Social Studies	
52	what is the name of the UCLA Center X Professional Learning Partner(s) (Coach(es)) you have been attending *Extended Day PD* or *All Day Pull-Out PD* with *this school year*?	Amparo Gonzalez-Soto	Jim Berger	Jon Kovach

Think about the *Extended Day PD* and *All Day Pull-Out PD* sessions offered *this school year* by your UCLA Center X Professional Learning Partner (Coach). Please circle your rating of the following:

		Excellent	Good	Fair	Poor	NA
53	The quality of materials received at the PD sessions.	E	G	F	P	NA
54	The engagement level and interest level of the PD sessions.	E	G	F	P	NA
55	The level of collaboration in the PD sessions.	E	G	F	P	NA
56	The usefulness of the PD sessions.	E	G	F	P	NA
57	The overall experience you have had with the PD sessions.	E	G	F	P	NA

Please Continue to the Next Page



UCLA Center X Teacher Semester Survey

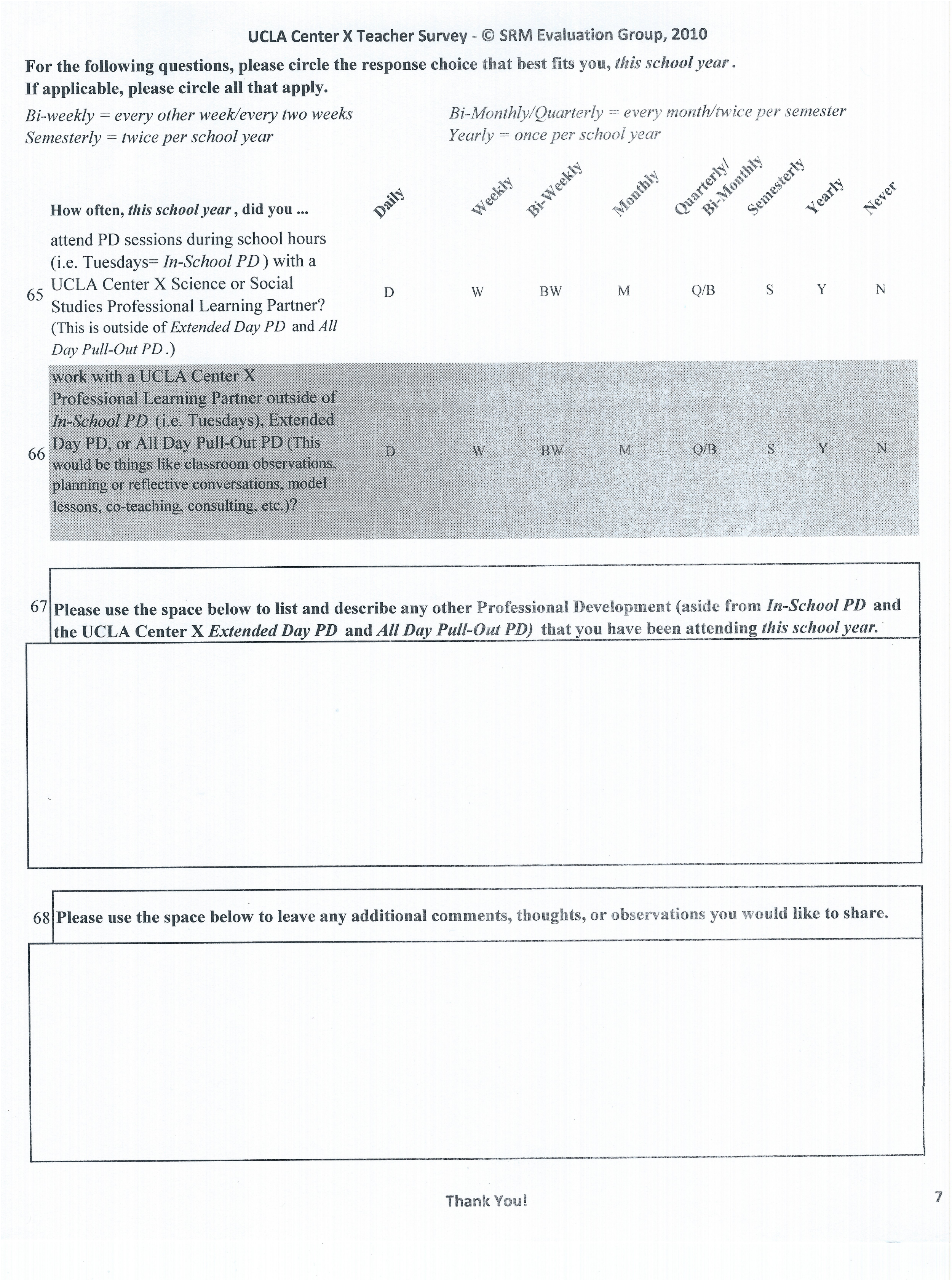
Please continue to think about *Extended Day PD* and *All Day Pull-Out PD* sessions offered this school year by your UCLA Center X Professional Learning Partner (Coach). Please circle how often the following occurred:

		Always	**Most of the Time**	**Some of the Time**	**Never**	**NA**
58	The goals/objectives of the PD sessions were clearly stated.	A	MT	ST	N	NA
59	The goals/objectives of the PD sessions were accomplished.	A	MT	ST	N	NA
60	The PD sessions specifically addressed the needs of Students with Disabilities.	A	MT	ST	N	NA
61	The PD sessions specifically addressed the needs of English Language Learners.	A	MT	ST	N	NA
62	The PD sessions specifically addressed the needs of African American students.	A	MT	ST	N	NA

63 **Please use the space below to describe in what ways attending PD sessions offered by UCLA Center X Professional Learning Partners (Coaches) has impacted you as a teacher.**

64 **Please use the space below to indicate in what ways Center X can improve the quality or usefulness of the *Extended Day PD* and/or *All Day Pull-Out PD* sessions offered by UCLA's Center X Professional Learning Partners (Coaches).**

Please Continue to the Next Page



UCLA Center X Teacher Survey - © SRM Evaluation Group, 2010

For the following questions, please circle the response choice that best fits you, *this school year*.
If applicable, please circle all that apply.

Bi-weekly = every other week/every two weeks *Bi-Monthly/Quarterly = every month/twice per semester*
Semesterly = twice per school year *Yearly = once per school year*

	How often, *this school year*, did you ...	**Daily**	**Weekly**	**Bi-Weekly**	**Monthly**	**Quarterly/ Bi-Monthly**	**Semesterly**	**Yearly**	**Never**
65	attend PD sessions during school hours (i.e. Tuesdays= *In-School PD*) with a UCLA Center X Science or Social Studies Professional Learning Partner? (This is outside of *Extended Day PD* and *All Day Pull-Out PD*.)	D	W	BW	M	Q/B	S	Y	N
66	work with a UCLA Center X Professional Learning Partner outside of *In-School PD* (i.e. Tuesdays), Extended Day PD, or All Day Pull-Out PD (This would be things like classroom observations, planning or reflective conversations, model lessons, co-teaching, consulting, etc.)?	D	W	BW	M	Q/B	S	Y	N

67 **Please use the space below to list and describe any other Professional Development (aside from *In-School PD* and the UCLA Center X *Extended Day PD* and *All Day Pull-Out PD)* that you have been attending *this school year.***

68 **Please use the space below to leave any additional comments, thoughts, or observations you would like to share.**

Thank You!

Appendix C
Construct Reliability Coefficients and Item Factor Loadings for Teacher Survey©

Construct	Item	EFA Factor Loadings	CFA Factor Loadings
Instructional Efficacy		(α = .855)	
	In general my classes are disciplined and well-behaved.[a]	0.883	0.814
	Students know that I expect hard work from them and they act accordingly.[a]	0.911	0.898
	For the most part, my students are engaged in my lessons.[a]	0.883	0.938
Collegiality		(α = .785)	
	Most of my colleagues share my beliefs and values about what the central mission of the school should be.[a]	0.767	0.743
	There is a great deal of cooperative effort among staff members here.[a]	0.804	0.799
	I can count on colleagues here when I feel discouraged about my teaching or my students.[a]	0.722	0.635
Leadership		(α = .673)	
	I participate in activities outside of the classroom for my school (i.e. student groups, school functions).[a]	0.588	0.669
	I am involved in decision-making for my school.[a]	0.782	0.802
	I initiate change in my school.[a]	0.794	0.803
	I have organized and presented professional development at my school.[b]	0.588	0.656
	I have organized and presented professional development at another school.[b]	0.338	0.429
Commitment to Ongoing Learning - Conferences		(α = .773)	
	In the past 12 months I have attended professional conferences.[c]	0.666	0.500
	In the past 12 months I have been the presenter at professional conferences.[c]	0.737	0.794
	In the next 12 months I plan to attend professional conferences.[c]	0.742	0.540
	In the next 12 months I plan to present at professional conferences.[c]	0.760	0.838
Reflective Practice - Videos		-	
	I have videotaped myself teaching science or social studies.[b]	-	0.852
	I have reviewed and reflected (alone or with colleagues) on a videotape of my teaching science or social studies.[b]	-	0.739
Collaboration		(α = .809)	
	I have collaborated with teachers within my department and my grade level.[b]	0.648	0.737
	I have collaborated with teachers within my department across grade levels.[b]	0.747	0.831
	I have collaborated with teachers across departments in my grade level.[b]	0.786	0.887
	I have collaborated with teachers across departments and across grade levels.[b]	0.740	0.881
	I have collaborated with teachers from a different school.[b]	0.488	0.427
	I have taught a common lesson created collaboratively within my department.[b]	0.559	0.613
	I have taught a common lesson created collaboratively across departments.[b]	0.588	0.678

Note. The SRM Evaluation Group conducted both an Exploratory Factor Analysis and a subsequent Confirmatory Factor Analysis on the constructs and items contained in the Teacher Survey. Data from six administrations across two programs were used to produce the final CFA factor loadings.

[a] Refers to items with the following scale: Strongly Disagree, Moderately Disagree, Neutral, Moderately Agree, Strongly Agree (1-5). [b] Refers to items with the following scale: Daily, Weekly, Bi-Weekly, Monthly, Quarterly/Bi-Monthly, Semesterly, Yearly, Never (1-8). [c] Refers to items with the following scale: 0 times, 1 time, 2 times, 3 times, 4 times, 5 times, 6 times or more (1-7).

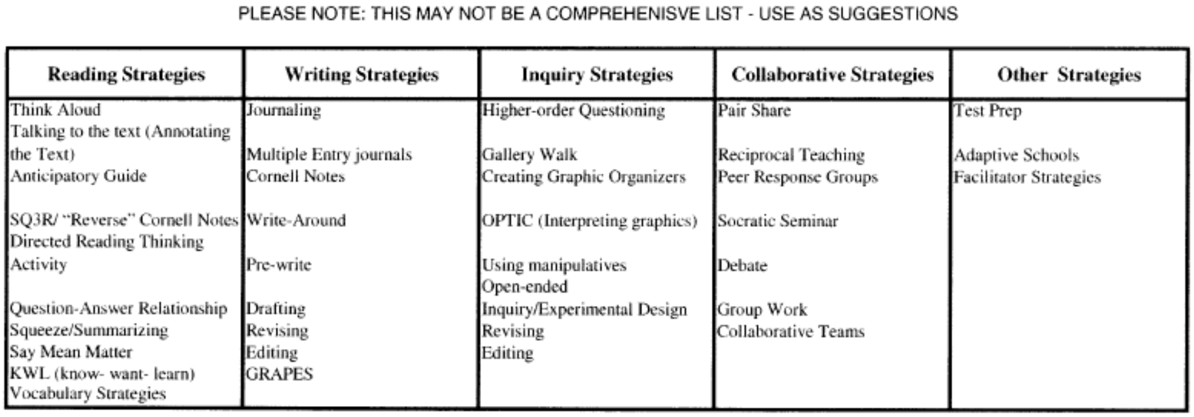
Appendix D
Expert Assessment
UCLA SRM Evaluation Group with UCLA Center X

School: _______________ Content Area: __________ PLP: _______ Date: __________

Literacy Instructional Strategy Assessment - Please use this form to indicate the frequency with which teachers implement the following sets of literacy instructional strategies. Please consider what you have observed during the 2011-2012 school year when indicating how often the following teachers implemented broad categories of literacy instructional strategies.

Implementation Scale:
1=Barely/Never, 2=Sometimes, 3=Often, 4=Most of the Time/Always

	NA	Reading Strategies				Writing Strategies				Inquiry Strategies				Collaborative Strategies				Other Strategies			
Teacher		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1																					
2																					
3																					
4																					
5																					
6																					
7																					
8																					
9																					
10																					
11																					
12																					
13																					
14																					



PLEASE NOTE: THIS MAY NOT BE A COMPREHENISVE LIST - USE AS SUGGESTIONS

Reading Strategies	Writing Strategies	Inquiry Strategies	Collaborative Strategies	Other Strategies
Think Aloud Talking to the text (Annotating the Text) Anticipatory Guide SQ3R/ "Reverse" Cornell Notes Directed Reading Thinking Activity Question-Answer Relationship Squeeze/Summarizing Say Mean Matter KWL (know- want- learn) Vocabulary Strategies	Journaling Multiple Entry journals Cornell Notes Write-Around Pre-write Drafting Revising Editing GRAPES	Higher-order Questioning Gallery Walk Creating Graphic Organizers OPTIC (Interpreting graphics) Using manipulatives Open-ended Inquiry/Experimental Design Revising Editing	Pair Share Reciprocal Teaching Peer Response Groups Socratic Seminar Debate Group Work Collaborative Teams	Test Prep Adaptive Schools Facilitator Strategies

Appendix E
Classroom Observation – Standards, Components, and Elements
UCLA SRM Evaluation Group and LAUSD T&LF

- Standard 2: Classroom Environment
  - Component 2A: Creating an Environment of Respect and Rapport
    - 2.A.1. Teacher Interactions with Students
    - 2.A.2 Student Interactions with One Another
    - 2.A.3. Classroom Climate
  - Component 2B: Establishing a Culture for Learning
    - 2.B.1. Importance of the Content
    - 2.B.2. Expectations for Learning and Achievement
    - 2.B.3. Student Ownership of their Work
    - 2.B.4. Physical Environment
  - Component 2C: Managing Classroom Procedures
    - 2.C.1. Management of Routines, Procedures, and Transitions
    - 2.C.2. Management of Materials and Supplies
    - 2.C.3. Performance of Non-Instructional Duties
    - 2.C.4. Management of Parent Leaders, other Volunteers and Paraprofessionals
  - Component 2D: Managing Student Behavior
    - 2.D.1. Expectations for Behavior
    - 2.D.2. Monitoring of Student Behavior
    - 2.D.3. Response to Student Behavior
- Standard 3: Instruction
  - Component 3A: Communicating with Students
    - 3.A.1. Expectations for Learning
    - 3.A.2. Directions and Procedures
    - 3.A.3. Explanations of Content
    - 3.A.4. Use of Academic Language
  - Component 3B: Using Questioning and Discussion Techniques
    - 3.B.1. Quality and Purpose of Questions
    - 3.B.2. Discussion Techniques
    - 3.B.3. Student Participation
  - Component 3C: Structures to Engage Students in Learning
    - 3.C.1. Standards-Based Projects, Activities, and Assignments
    - 3.C.2. Purposeful and Productive Instructional Groups
    - 3.C.3. Use of Available Instructional Materials, Technology and Resources
    - 3.C.4. Structure and Pacing
  - Component 3D: Delivery of Instruction
    - 3.D.1. Assessment Criteria
    - 3.D.2. Monitoring of Student Learning
    - 3.D.3. Feedback to Students
    - 3.D.4. Student Self-Assessment and Monitoring of Progress
  - Component 3E: Demonstrating Flexibility and Responsiveness
    - 3.E.1. Responds and Adjusts to Meet Student Needs
    - 3.E.2. Persistence

Appendix F
Testing of Assumptions in Regards to Analysis for Research Question 1

In this, and the following appendix, I address the assumptions for the statistical procedures conducted to answer research questions 1 and 2. Given the small sample size, the ability to fully reject statistical assumptions is limited. Minor violations are addressed by modifying statistical procedures, when appropriate. Results from checking assumptions are arranged by research question, instrument, statistical procedure, and time.

Teacher Survey

A multivariate analysis of variance (one-way MANOVA) is conducted at T1, T2, and T3 to better understand group differences in mean scores on survey constructs (instructional efficacy, collegiality, leadership, ongoing learning, reflective practice, and collaboration) (reflective practice data is only available at T2 and T3). I check to see if the following assumptions are met: absence of univariate and multivariate outliers; presence of multivariate normality; linear relationship between independent variable groups and the dependent variables; homogeneity of variance-covariance matrices, and absence of multicollinearity. I address each set of assumptions by time of survey administration.

T1 – Inspection of boxplots for values greater than 1.5 box-lengths from the edge of the box reveals two univariate outliers in the low participation group and 3 univariate outliers in the high participation group. Closer inspection of these 5 cases indicates all values are reasonable and so they are kept in the analysis. There are no multivariate outliers in the data, as assessed by Mahalanobis distance ($p > .001$). Shapiro-Wilk tests of normality indicate the 5 dependent variables are largely normally distributed for all groups at T1 ($p > .05$). In only 2 instances is $p <$

.05, both times in the moderate group, for collegiality and leadership. Overall, only slight violations of normality are identified. Linear relationships between dependent variables are evident, as assessed by scatterplots. A Pearson correlation table shows no more than a moderate correlation between variables, indicating no multicollinearity. There is homogeneity of variance-covariance matrices, as assessed by Box's test of equality of covariance matrices (p = .519). There are homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance for 4/5 variables (p > .05). For instructional efficacy, p = .007 indicating unequal variances for that construct. To address the equal variances assumption violation, I accept a lower level (more stringent) of statistical significance for the MANOVA result.

T2 – Inspection of boxplots for values greater than 1.5 box-lengths from the edge of the box reveals several univariate outliers in each of the participation groups. Close inspection of each of these cases indicates all values are reasonable and so they are kept in the analysis. There are no multivariate outliers in the data, as assessed by Mahalanobis distance (p > .001). Shapiro-Wilk tests of normality indicate that 2 of the dependent variables (leadership and collaboration) are normally distributed for each participation group (p > .05). The other 4 dependent variables are not normally distributed for all groups at T2 (p < .05). The MANOVA is relatively robust to normality threats. Linear relationships between dependent variables are evident, as assessed by scatterplots. A Pearson correlation table shows no more than a moderate correlation between variables, indicating no multicollinearity. There is questionable homogeneity of variance-covariance matrices, as assessed by Box's test of equality of covariance matrices (p = .001). There are homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance for 4/6 variables (p > .05). For instructional efficacy, p = .037, and for collegiality, p = .022, indicating unequal variances for those constructs. To address threats to equal covariance and variances, I accept a lower level of statistical significance for the MANOVA result.

T3 – Inspection of boxplots for values greater than 1.5 box-lengths from the edge of the box reveals several univariate outliers in each of the participation groups. Close inspection of each of these cases indicates all values are reasonable and so they are kept in the analysis. There are no multivariate outliers in the data, as assessed by Mahalanobis distance ($p > .001$). Shapiro-Wilk tests of normality indicate that 3 of the dependent variables (leadership, collegiality, and collaboration) are normally distributed for each participation group ($p > .05$). The other 3 dependent variables are not normally distributed for all groups at T3 ($p < .05$). The one-way MANOVA is fairly robust to such minor deviations from normality. Linear relationships between dependent variables are evident, as assessed by scatterplots. A Pearson correlation table shows no more than a moderate correlation between variables, indicating no multicollinearity. There is homogeneity of variance-covariance matrices, as assessed by Box's test of equality of covariance matrices ($p = .008$). There are homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance for 5/6 variables ($p > .05$). For ongoing learning, $p = .014$, indicating unequal variances for that construct. To address the equal variance assumption violation, I accept a lower level of statistical significance for the MANOVA result by using a Bonferroni correction (to reduce Type I error) (p < .025 (*) rather than p < .05) when interpreting follow-up univariate ANOVAs, and use the Games-Howell post-hoc test as the multiple comparison procedure.

A series of repeated measures ANOVAs are conducted for the low, moderate, and high participation groups to test for significant change over time on each of the teacher effectiveness constructs housed within the teacher survey. In addition to the above tested assumptions, the assumption of sphericity is tested with Mauchly's test for Sphericity – results are organized by participation group.

Low Participation Group – Mauchly's test for Sphericity indicates the following for each construct: instructional efficacy – the assumption of sphericity is met, $X^2(2) = 4.980$, $p = .083$; collegiality – the assumption of sphericity is met, $X^2(2) = 3.195$, $p = .202$; leadership – the assumption of sphericity is met, $X^2(2) = 2.027$, $p = .363$; ongoing learning – the assumption of sphericity is met, $X^2(2) = .630$, $p = .730$; and for collaboration – the assumption of sphericity is met, $X^2(2) = 4.404$, $p = .111$.

Moderate Participation Group – Mauchly's test for Sphericity indicates the following for each construct: instructional efficacy – the assumption of sphericity is met, $X^2(2) = 3.719$, $p = .156$; collegiality – the assumption of sphericity is met, $X^2(2) = 1.726$, $p = .422$; leadership – the assumption of sphericity is met, $X^2(2) = .184$, $p = .912$; ongoing learning – the assumption of sphericity is met, $X^2(2) = 1.348$, $p = .510$; and for collaboration – the assumption of sphericity is met, $X^2(2) = 2.733$, $p = .255$.

High Participation Group – Mauchly's Test for Sphericity indicates the following for each construct: instructional efficacy – the assumption of sphericity is met, $X^2(2) = 2.135$, $p = .334$; collegiality – the assumption of sphericity is met, $X^2(2) = .537$, $p = .765$; leadership – the assumption of sphericity is met, $X^2(2) = 1.251$, $p = .535$; ongoing learning – the assumption of sphericity is met, $X^2(2) = .595$, $p = .74$; and for collaboration – the assumption of sphericity is met, $X^2(2) = 1.141$, $p = .565$.

Expert Assessment

There are three time points (Y2T2, Y3T1, Y3T2) available for analysis with the expert assessment data. A mixed ANOVA is run to determine whether there are differences between participation groups (low, moderate, high) over time (T1, T2, T3) on their overall strategy use (DV). For this analysis, there is data available at three time points for 20 teachers in the no/low participation group, 16 teachers in the moderate participation group, and 12 teachers in the high participation group. 2 outliers (low scores) are detected in the high participation group for the dependent variable overall strategy use (as assessed by inspection of a boxplot for values greater than 1.5 box-lengths from the edge of the box). Close inspection of these two points reveals that they are well within reason and they are kept in the analysis (studentized residuals are all under ±3 supporting this decision).

The Shapiro-Wilk tests of normality reveal that there are in fact issues with normality. For 7 of the 9 tests, overall strategy use is not normally distributed, $p < .05$. In order to address the violations of normality, I attempt to transform the data at the three time points. For the time 1 and time 2 variable, I use the transformation procedure for moderately positively skewed data (sqrt(DV)). At time 3 I use the transformation procedure for moderately negatively skewed data (sqrt(5-DV)). Although skewness statistics were generally within the moderate range, for a few places, skewness statistics indicted a more heavily skewed (strong) distribution. The difficulty in selecting an appropriate transformation procedure lies within the fact that the strength and direction of the skewed distribution does not remain the same for each of the groups or for each time point. As expected, higher participation groups have a greater concentration of high ratings and lower participation groups have a greater concentration of low ratings. Because of this, none of the transformations adequately address the violations of normality for each category of the independent variable (group). Data remains moderately skewed (6/9 tests). Due to the relatively robust nature of the mixed ANOVA to deviations from normality, analysis proceeds with the original variables. Inspection of Normal Q-Q Plots (of studentized residuals) for each level of the within-subjects factor (time), are not too distorted from the diagonal line, supporting the use of original variables.

Levene's Test of Homogeneity of Variance indicates the assumption of equal variances is met ($p > .05$). There is also homogeneity of covariance, as assessed by Box's Test of Equality of Covariance Matrices ($p = .089$). Mauchly's Test of Sphericity is statistically significant ($p < .05$) and therefore I use the Huynh-Feldt correction (epsilon > .75) to interpret the F-ratio in the Tests of Within-Subjects Effects (Field, 2012).

Classroom Observations

For 16 teachers, I am able to conduct analyses looking at change over time. Given that the program expects changes in ratings over time for these folks, especially for the high-participation group, I look at change over time separately for these two groups. To test for change over time, a series of paired sample t-tests are run on classroom environment (including 4 sub-constructs), instruction (including 5 sub-constructs), and total strategy usage (including 5 sub-constructs). For the moderate participation group, across the 17 constructs and sub-constructs, only four outliers are identified (across 3 teachers); 2 more than 1.5 box-lengths from the edge of the box in the boxplot, and 2 more than 3 box-lengths from the edge. Outlier mean difference values are equally split between, laying above, and laying below the boxplots. Close inspection of these values does not reveal them to be extreme and they are kept in the analysis. The Shapiro-Wilk test of normality reveals only one variable for which difference scores are violating a normal distribution. The difference scores for the collaborative strategies-Y3T2 and collaborative strategies-Y3T1 scores are not normally distributed, *p* = .021.

For the high-level participation group, across the 17 constructs and sub-constructs, 12 distinct points are identified as outliers, all over 1.5 lengths from the edge of the box in the boxplot. Outlier mean difference values are proportionately split between, laying above, andlaying below the boxplots. Close inspection of these 12 values (across 8 teachers) reveals that none are extreme and therefore they are all kept in the analysis. The Shapiro-Wilk test of normality reveals 5 variables for which there is a violation of normality between difference scores (Y3T2-Y3T1): S2CD (*p* = .033); S3CA (*p* = .037); reading strategies (*p* = .001); collaborative strategies (*p* = .002); and other strategies (*p* = .025).

A series of one-way ANOVAs are conducted on classroom environment and instruction to test for participation group differences. In examining the data for outliers (by way of boxplots), only 2 data points for the no/low participation group are identified and 3 data points for the moderate participation group are identified. Close examination of these points reveals that they are not extreme and they are kept in the analysis. Scores on constructs and sub-constructs are largely normally distributed, as assessed by Shapiro-Wilk's test of normality (*p* > .05 for 31 out of 33 instances). One-way ANOVAs are relatively robust to minor violations of normality.

A one-way MANOVA is conducted on average total strategy usage for low, moderate, and high participation groups. Five measures of strategy usage are assessed: reading, writing, inquiry, collaborative, and other strategies. Preliminary checking of assumptions reveals that the data is normally distributed, as assessed by Shapiro-Wilk test of normality (*p* > .05); there are no univariate or multivariate outliers, as assessed by boxplots and Manhalanobis distance (*p* > .001), respectively; there are linear relationships between variables, as assessed by scatterplots; no multicollinearity (moderate correlations between all variables: r < .60); there is homogeneity of variance-covariance matrices, as assessed by Box's M test (*p* = .447); and there is homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance (*p* > .05).

Appendix G
Testing of Assumptions in Regards to Analysis for Research Question 2

Pearson correlation coefficients are produced to explore relationships between the 11 continuous variables (10 major teacher effectiveness constructs and APD) used in the study. Inspection of scatterplots reveals linear relationships between variables. Outliers for each teacher effectiveness construct have already been identified in the checking of assumptions for RQ 1. All points have been inspected and prove to be legitimate. Likewise, Shapiro-Wilk's tests of normality, produced for RQ 1 reveal constructs for which there are slight violations of normality. Given the small sample size, minor violations are not worrisome.

For all regression models, assumptions are tested (independence of cases, linearity, and homoscedasticity) and adequately satisfied. For all models, there is independence of residuals, as assessed by a Durbin-Watson statistic near 2 (actual range is from .77 to 2.13, with the majority being around 1.8). The assumptions of linearity and homoscedasticity are met, as assessed by scatterplots. Inspection of collinearity statistics reveals that there are no issues with multicollinearity with the independent variables (Tolerance values are > .1). Standardized residuals and studentized deleted residuals are all less than ±3 indicating there are no problems with outliers that may be detrimental to the fit of the regression equations. Leverage values are generally regarded as safe (< .02) or potentially risky (< .05, the highest being .03) and are not of concern. Cook's Distance values are all under 1 indicating the absence of influential cases. Standardized residuals appear to be normally distributed, as conferred with histograms, P-P Plots, and Q-Q Plots, indicating normality is not a problem. Being that all assumptions are adequately addressed multiple regression models are produced to better understand relationships between variables, including predictive power.

References

Aguerrebere, J. A. (2010). *Letter to Arne Duncan: Secretary's priorities for discretionary grant programs* (Docket ID: ED-OS-2010-0011). Retrieved from http://www.nbpts.org/

Alwin, D. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken, NJ: Wiley.

Armstrong, J. & Anthes, K. (2001). How data can help: Putting information to work to raise student achievement. *American School Board Journal, 188*(11), 38-41.

Bakke, P. A. (1999). *Perceptions of characteristics of effective teachers*. Retrieved from http://ucla.worldcat.org/

Baker, E.L. et al. (2010). *Problems with the use of student test scores to evaluation teachers* (EPI Briefing Paper 278). Washington, D.C.: Economic Policy Institute.

Bang-Jensen, V. (1986). The View from Next Door: A Look at peer supervision. In K. K. Zumwalt, (Ed.), *Improving Teaching: 1986 ASCD Yearbook*. Alexandria, VA: Association for Supervision and Curriculum Development.

Bausell, R. B. (2011). A new measure for classroom quality. *The New York Times*. Retrieved from http://www.nytimes.com/2011/05/01/opinion/01bausell.html?_r=2

Biemer, P. (2004). *Measurement error in surveys*. Hoboken, NJ: Wiley.

Biemer, P., & Lyberg, L. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.

Bill & Melinda Gates Foundation. (2010). *MET project research paper: Learning about teaching: Initial findings from the Measures of Effective Teaching project*. Retrieved from www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Bill & Melinda Gates Foundation. (2013). *Feedback for better teaching: Nine principles for using measures of effective teaching*. Retrieved from http://www.metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principles%20Paper.pdf

Birman, B. F., Desimone, L., Porter, A.C., & Garet, M.S. (2000). Designing professional development that works. *Educational Leadership*. Retrieved from www.ascd.org/.../educational_leadership/.../Designing_Professional_Development_That_Works.aspx

Borko, H., Stecher, B. M., Martinez, F., Kuffner, K.L., Barnes, D., Arnold, S. C., et al. (2006). *Using Classroom Artifacts to Measure Instructional Practice in Middle School Science: A Two-State Field Test* (CSE Technical Report 690). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Bradburn, N., Sudman, S., & Wasink, B. (2004). *Asking questions: The definitive guide to questionnaire design – For market research, political polls, and social and health questionnaires*. San Francisco, CA: Johnson Wiley & Sons.

Brown, S. L. (2004). *High school students' perceptions of teacher effectiveness: Student ratings and teacher reflections*. Retrieved from http://ucla.worldcat.org/

Buddin, R. (2010). How effective are Los Angeles Elementary Teachers and Schools? *LA Times Electronic Report*. Retrieved from http://www.latimes.com/

California Department of Education. (2013). DataQuest. Retrieved from http://data1.cde.ca.gov/dataquest

California P-16 Council. *Closing the Achievement Gap*. Retrieved from http://www.cde.ca.gov/eo/in/pc/documents/yr08ctagrpt0122.pdf

California Post Secondary Education Commission (CPEC). (2005). *Improving teacher quality Program*. Retrieved from http://www.cpec.ca.gov

California Post Secondary Education Commission (CPEC). (2009). *Teacher professional development in California: A status of review.* Retrieved from http://www.cpec.ca.gov

California Post Secondary Education Commission (CPEC): Website: http://www.cpec.ca.gov

California State Board of Education. (2006). *California's revised state plan for No Child Left Behind: Highly qualified teacher*. Retrieved from www.2.ed.gov/programs/teacherqual/hqtplans/ca.doc

Carter, P. J. (2008). *Defining teacher quality: An examination of the relationship between measures of teachers' instructional behaviors and measures of their students' academic progress*. Retrieved from http://ucla.worldcat.org/

Chait, R. (2009). *Ensuring effective teachers for all students*. Center for American Progress. Retrieved from http://www.americanprogress.org/issues/2009/05/pdf/teacher_effectiveness.pdf

Chamberlain, G. (2013). Predictive effects of teachers and school on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences (PNAS), 110*(43), 17176-17182. doi: 10.1073/pnas.1315746110

Clotfelter, Ladd, & Vigdor. (2008). *What do we know about the relationship between student achievement and teachers' educational attainment and experience, which is the traditional way that teacher salaries are determined?* The Center for Educator Compensation Reform. Retrieved from http://cecr.ed.gov/guides/researchSyntheses/Research%20Synthesis_Q%20A2.pdf

Cohen, D., & Hill, H. (1998). *Instructional policy and classroom performance: The mathematics reform in California*. (RR-39). Philadelphia, PA: Consortium for Policy Research in Education.

Cohen, D., McLaughlin, M., & Talbert, J. (Eds.). (1993). *Teaching for understanding: Challenges for policy and practice.* San Francisco: Josey-Bass.

Commission on Teacher Credentialing. (2009). *California standards for the teaching profession (CSTP).* Retrieved from http://www.ctc.ca.gov/educator-prep/standards/CSTP-2009.pdf

Comprehensive Center for Teacher Quality (U.S.). (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Corcoran, T. (1995). *Helping teachers teach well: Transforming professional development*. (CPRE Policy Brief). Philadelphia, PA: Consortium for Policy Research in Education.

Corcoran, T, Shields, P., & Zucker, A. (1998). *The SSIs and professional development for teachers: Evaluation of the National Science Foundation's statewide systemic initiatives (SSI) program*. Menlo Park, CA: SRI International.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice.* Princeton, NJ: Educational Testing Service.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (1993). Reframing the school reform agenda: Developing capacity for school transformation. *Phi Delta Kappa, 74*, 735–761.

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching*. For: The National Commission on Teaching and America's Future.

Darling-Hammond, L. (2000). *Studies of excellence in teacher education (3 volumes).* Washington, DC: American Association of Colleges for Teacher Education.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved from http://epaa.asu.edu/epaa/v8n1/

Darling-Hammond, L. (2001). *Developing and assessing teacher effectiveness: Real reform*. Presentation given at the California Postsecondary Education Commission's Improving Teacher Quality Program Annual Project Director's Meeting in Los Angeles, California. Retrieved from http://www.cpec.ca.gov/FederalPrograms/DevelopingAndAssessing TeacherEffectiveness.pdf

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching*. Retrieved from http://www.americanprogress.org

District of Columbia Public Schools: Website: http://www.dcps.dc.gov/portal/site/DCPS/

DuFour, R. (2004). What is a "professional learning community"? *Educational Leadershi*p, *61*(8), 6–11.

Earl, L. M. & Katz, S. (2006). *Leading Schools in a Data-Rich World: Harnessing Data for School Improvement.* Thousand Oaks, CA. Corwin Press.

Easton, L.B. (2004). *Powerful designs for professional learning*. Oxford, Ohio: National Staff Development Council.

Easton, L.B. (2008). From professional development to professional learning. *Phi Delta Kappan, 89*(10), 755-759.

Edwards, J. M. (2005). *No Child Left Behind teacher quality policies, practices, and teacher effectiveness research*. Retrieved from http://ucla.worldcat.org/

Educators 4 Excellence. (2012). *Breaking the stalemate: LA teachers take on evaluation*. Retrieved from http://www.unitedwayla.org/wp-content/uploads/2012/06/E4E-LA-Teacher_Evaluation-final1.pdf

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" education research. *Educational Researcher*, *32*(7), 31–38.

Elmore, R. (2002). *Bridging the gap between standards and achievement: The imperative for professional development in education*. Albert Shanker Institute. Retrieved from http://www.shankerinstitute.org/Downloads/Bridging_Gap.pdf

Erickson, F., & Gutierrez, K. (2002). Culture, rigor, and science in educational research. *Educational Researcher*, *31*(8), 21–24.

Fabiano, L. (1999). *Measuring teacher qualifications.* Report prepared for U.S. Department of Education, Office of Educational Research and Improvement, and National Center for Education Statistics. Retrieved from nces.ed.gov/pubs99/199904.pdf

Feldman, J. & Tung, R. (2001). Using data-based inquiry and decision making to improve instruction. *ERS Spectrum, 19*(3)*,* 10-19.

Ferguson, R.F. (1999). Teachers' perceptions and expectations and the black-white test score gap. In C. Jencks, & M. Phillips (Eds.), *The black-white test score gap and can schools narrow the black-white test score gap?* (pp. 273-374). Washington, DC: Brookings Institution Press.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher*, *31*(8), 4–14.

Field, A. (2012). *Discovering Statistics*. Retrieved from http://www.discoveringstatistics.com

Florida Department of Education. (2010). *Florida Department of Education professional development system evaluation protocol*. Bureau of Educator Recruitment, Development and Retention.

Forte, B. (1999) *The impact of professional development on teacher performance and student achievement*. Retrieved from http://ucla.worldcat.org/

Fowler, F. (1995). Improving survey questions: Design and evaluation. Thousand Oaks, CA: Sage Publications.

Gage, N.L. (1972). *Teacher effectiveness and teacher education*. Palo Alto, CA: Pacific Books, Publishers.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Education Research Journal, 38*(4), 915-945.

Gargani+ Company. (2009). *Final evaluation report: Cluster-Randomized trials of simultaneously implemented professional development programs for history and science teachers*. A report prepared for the California Postsecondary Education Commission (CPEC).

Gilman, A. G., Emhuff, J., & Hamm, J. *Improving teacher attitude and morale through maintaining teacher effectiveness: An Indiana staff development model.* Mount Vernon School District, IN.; Indiana State Univ., Terre Haute. Professional School Services.

Gloudemans, P. (Ed.). (2011). Effective teaching as a civil right. *Voices in Urban Education, 31*. Annenberg Institute for School Reform.

Goe, L. (2010). Evaluating teacher effectiveness: An overview. *LAUSD Convocation on Educator Effectiveness*. California: Los Angeles.

Goe, L., Bell, C., & Little, O. (2008). Ap*proaches to evaluating teacher effectiveness: A research synthesis*. National Comprehensive Center for Teacher Quality. Retrieved from http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf

Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Retrieved from http://www.nbpts.org/

Gordon, R., Kane, T.J., & Staiger, D.O. (2009). *Identifying effective teachers using performance on the job.* The Brookings Institution. Retrieved from http://www.brookings.edu/~/media/Files/rc/papers/2006/04education_gordon/200604hamilton_1.pdf

Graham, P. (2007). Improving teacher effectiveness through structured collaboration: A case study of a professional learning community. *RMLE Online: Research in Middle Level Education, 31*(1), 1-17.

Guskey, T. R. (2000). *Evaluating Professional Development*. Thousand Oaks, CA: Corwin Press, Inc.

Guskey, T.R. (2002). Does it make a difference? Evaluating professional development. Educational Leadership. Retrieved from www.ascd.org/.../educational_leadership/.../Does_It_Make_a_Difference¢_Evaluating_Professional_Development.aspx

Guskey, T.R. (2003). Analyzing lists of the characteristics of effective professional development to promote visionary leadership. *NASSP Bulletin 87*(4), 4-20. doi: 10.1177/019263650308763702

Guskey, T.R. (2003). What makes professional development effective? *The Phi Delta Kappan, 84*(10), 748-750. Retrieved from http://www.jstor.org/stable/20440475

Guskey, T.R. (2006). A conversation with Thomas R. Guskey. In Krieder, H. (Ed.), *The Evaluation Exchange XI*(4), 5. Harvard Family Research Project, Harvard Graduate School of Education.

Guskey, T.R., & Yoon, K.S. (2009). What works in professional development? *The Phi Delta Kappan, 90*(7), 495-500. Retrieved from http://www.pdkintl.org/search.htm.

Hanushek, E. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*(1), 84-117.

Harris, A., & Muijs, D. (2005). *Improving schools through teacher leadership*. Maidenhead: Open University Press.

Haycock, K. (1998). Good teaching matters…a lot. *Thinking K-163,* No. 2.

Hill, C. L. (2002). *What are the characteristics of effective teaching? A comparative study of stakeholder perceptions*. Retrieved from http://ucla.worldcat.org/

Hoy, W.K., & Woolfolk, A.E. (1993). Teachers' sense of efficacy and the organizational health of schools. *Elementary School Journal, 93*, 355-372.

Kemp, & Hall. (1992). *Teacher effectiveness*. North Central Regional Educational Laboratory.

Kennedy, M.M. (1998). Form and substance in inservice teachers education. Research *Monograph No. 13*. Madison, Wis: National Institute for Science Education, University of Wisconsin-Madison.

Killion, J. (2006). Evaluating the impact of professional development in eight steps. In Krieder, H. (Ed.), *The Evaluation Exchange XI*(4), 5. Harvard Family Research Project, Harvard Graduate School of Education.

Kim, J. S. & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 34*(8), 3-13.

Klonsky, M. (2002). Small schools and teacher professional development. *ERIC Digest.* ED470949

Junker, B., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., et al. (2006). *Overview of the instructional quality assessment* (CSE Technical Report No. 671). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lansman, R. R. (2006). *A case study of teacher evaluation and supervision at a high-achieving urban elementary school*. Retrieved from http://ucla.worldcat.org/

Learning Point Associates, (2006). Using data as a school improvement tool. Published by *North Central Regional Educational Laboratory.* Naperville, IL.

Learning Forward. (2011). *Learning Forward's standards for professional learning*. Retrieved from http://www.learningforward.org/standards/#.UCGHrsie44Q

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Linn, R.L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives*, vol. 13, no. 33.

Little, O., Goe, L., & Bell, C. (2101). A practical guide to evaluating teacher effectiveness. *LAUSD Convocation on Educator Effectiveness*. California: Los Angeles.

Los Angeles Times. (2013). *Mapping L.A. Neighborhoods.* Retrieved from http://projects.latimes.com/mapping-la/neighborhoods/

Los Angeles Unified School District (LAUSD): Website: http://etf.lausd.net/

Los Angeles Unified School District. (2010).*Building our certificated employee development system: Three-year strategic plan*. Retrieved from http://etf.lausd.net/

Los Angeles Unified School District. (2010). *Teacher Effectiveness Task Force: Final report.* Retrieved from http://etf.lausd.net/

Los Angeles Unified School District. (2010). *Teacher Effectiveness Task Force: Initial draft recommendations for revamping the teacher evaluation system*. Retrieved from http://etf.lausd.net/

Los Angeles Unified School District. (2011). *Information Brief: LAUSD's Teaching and Learning Framework*. Retrieved from http://talentmanagement.lausd.net

Los Angeles Unified School District (2012). *My Data* (Data file). Retrieved from http://notebook.lausd.net/portal/page?_pageid=33,1047597&_dad=ptl&_schema=PTL_EP.

Loucks-Horsley, S., Hewson, P.W., Love, N., & Stiles, K.E. (1998). Designing professional development for teachers of science and mathematics. A project of the National Institute for Science Education. Thousand Oaks, CA: Corwin Press.

Lustick, D., & Sykes, G. (2006). *National Board Certification as professional development: What are teachers learning?* Retrieved from http://www.nbpts.org/

MacCalla, N. M., Dillman, L., & Alkin, M. (2013). *Final evaluation report: A quasi-experimental design study of a professional development program for middle school science and social studies teachers.* Report prepared by: UCLA SRM Evaluation Group for the California Department of Education (CDE).

McLaughlin, M.W., & Marsh, D.D. (1978). Staff development and school change. *Teachers College Record, 80,* 70-94.

Mintrop, Heinrich, & Trujillo. (2007). The Practical Relevance of Accountability Systems for School Improvement: A Descriptive Analysis of California Schools. *Educational Evaluation and Policy Analysis*, *29*(4), 319-352.

Moss, P.A. (1994). Can there be validity without reliability? *Educational Research, 23*(2), 5-12.

Moss, P.A., Schutz, A.M., & Collins, K.M. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education, 12*(2), 139-161.

Murnane, R.J. (1985). *Do effective teachers have common characteristics: Interpreting the quantitative research evidence*. Paper presented at the National Research Council Conference on Teacher Quality in Science and Mathematics. Washington, D.C.

National Board for Professional Teaching Standards: Website: http://www.nbpts.org/

National Board for Professional Teaching Standards. (2002). *What teachers should know and be able to do*. Retrieved from http://www.nbpts.org/

National Board for Professional Teaching Standards. (2003). *NBPTS early adolescence science standards: For teachers of students ages 11-15 (Second Edition)*. Retrieved from http://www.nbpts.org/

National Board for Professional Teaching Standards. (2007). *A research guide on National Board Certification® of teachers*. Retrieved from http://www.nbpts.org/

National Board for Professional Teaching Standards. (2009).*Early adolescence: Science: scoring guide for candidates*. Retrieved from http://www.nbpts.org/userfiles/File/EA_Science_Scoring_Guide.pdf

National Board for Professional Teaching Standards. (2010). *EA/Science portfolio entry directions*. Retrieved from http://www.nbpts.org/

National Board for Professional Teaching Standards. (2010). *Social Studies-History standards: For teachers of students ages 7-18+ (Second Edition)*. Retrieved from http://www.nbpts.org/

National Board for Professional Teaching Standards. (2010). *Part 1: Understanding and interpreting your scores*. Pearson. Retrieved from http://www.nbpts.org/userfiles/File/Part1_Interpreting_your_score.pdf

National Board for Professional Teaching Standards. (2011). *EA/Social Studies-History portfolio entry directions*. Retrieved from http://www.nbpts.org/

National Board for Professional Teaching Standards. (2011). *Early adolescence: Social Studies-History: Scoring guide for candidates*. Retrieved from http://www.nbpts.org/userfiles/file/EA_SSH_Scoring_Guide_041811.pdf

National Board for Professional Teaching Standards. (2011). *Evaluation of evidence guide: Early adolescence/science*. Retrieved from http://www.nbpts.org/userfiles/File/EA_Science_EvalEvidence_Guide.pdf

National Board for Professional Teaching Standards. (2011). *Evaluation of evidence guide: Early adolescence/social studies-history*. Retrieved from http://www.nbpts.org/userfiles/file/EA_SSH_EvalEvidence_Guide_041211.pdf

National Center for Educational Statistics. (2008). *Revenues and expenditures for public elementary and secondary education: School year 2005-2006 (Fiscal year 2006).* Washington, D.C., Institute for Educational Sciences, U.S. Department of Education. Retrieved from http://www.nces.ed.gov/pubs2008/expenditures.

National Council on Teacher Quality. (2011). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies*. Retrieved from http://www.nctq.org/dmsView/State_of_the_States_Teacher_Evaluation_and_Effectiveness_Policies_NCTQ_Report

National Council on Teacher Quality. (2011). *Teacher quality roadmap: Improving policies and practices in LAUSD*. Retrieved from http://www.nctq.org/dmsView/Teacher_Quality_Roadmap_Improving_Policies_and_Practices_in_LAUSD_NCTQ_Report

National Council on Teacher Quality. (2012). *State of the states 2012: Teacher effectiveness policies*. Retrieved from http://www.nctq.org/dmsView/State_of_the_States_2012_Teacher_Effectiveness_Policies_NCTQ_Report

National Education Association website: htpp://www.nea.org

National Research Council (2002). *Scientific research in education*. Washington, DC: National Academy Press.

National Staff Development Council (NSDC). (2001). *NSDC's standards for staff development.* Oxford, Ohio: National Staff Development Council.

Oakes, J., Franke, M. L., Quartz, K. H., & Rogers, J. (2002). Research for High-Quality Urban Teaching: Defining It, Developing It, Assessing It. *Journal of Teacher Education, 53(3),* 228-34.

Pelligrino, J. W. & Goldman, S. R. (2002). Be careful what you wish for – you may get it: Educational research in the spotlight. *Educational Researcher*, *31*(8), 15-17.

Phillips, M., & Strunk, K. (2010). Using research to support effective teaching. *LAUSD Convocation on Educator Effectiveness*. California: Los Angeles.

Presser, S., Couper, M., Lessler, J., Martin, E., Martin, J., Rothgeb, J., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*(1), 109-130. doi: 10.1093/poq/nfh008

Puma, M., & Raphael, J. (2001). *Evaluating standards-based professional development for teachers: A handbook for practitioners*. Report prepared by The Urban Institute for the U.S. Department of Education.

Quartz, K. (2007). Annual survey of California science project participants. Report prepared for UCLA Center X.

Rice, J.K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes*. Washington, DC: Economic Policy Institute.

Rivers, J.C., & Saunders, W. (2002). *Teacher quality and equity in educational opportunity: Findings and policy implications in teacher quality*. Stanford, CA: Hoover Institution Press.

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

Russo, M. (2006). Teacher professional development: How do we establish it and know that it's working? In Krieder, H. (Ed.), *The Evaluation Exchange XI*(4), 6-7. Harvard Family Research Project, Harvard Graduate School of Education.

Sanders, W.L., & Horn, S.P. (1994). The Tennessee value-added assessment system (TVASS): Mixed methodology in educational assessment. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Sartain, L.(2010). Measuring teaching practice: Lessons from Chicago public schools. *LAUSD Convocation on Educator Effectiveness*. California: Los Angeles.

Saunders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville: Value Added Research and Assessment Center, University of Tennessee.

Schon, D. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Schulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Schweig, J., Applegate, M., & Quartz, K.H. (2010). UCLA IMPACT: Integrating research and practice using multiple measures. In Quartz, K., & Applegate, M. (Eds.), Multiple Measures of Good Teaching. (2011, Spring). *Center XChange*. UC Regents. Retrieved from http://centerx.gseis.ucla.edu/xchange-repository/spring-2011

State Superintendent of Public Instruction Tom Torlakson's Task Force on Educator Excellence. (2012). *Greatness by Design: Supporting Outstanding Teaching to Sustain a Golden State*. Retrieved from http://www.cde.ca.gov/eo/in/documents/greatnessfinal.pdf

Stronge, J. H. (2010). *Effective teachers = student achievement: What the research says.* New York, NY: Taylor & Francis.

Sudman, Bradburn, S., & Bradburn, N. (1982). Asking questions: A practical guide to questionnaire design. San Francisco, CA: Jossey-Bass.

Sutherland, S. (2004). Creating a culture of data use for continuous improvement: A case study of an edison project school. *American Journal of Evaluation, 25*(3), 277-293.

Taylor, P., & Walpole, A. (1999). *Teacher effectiveness*. North Central Regional Educational Laboratory.

Tennessee Department of Education: Website: http://www.team-tn.org

The New Teacher Project. (2009, September). *"Interpreting Race to the Top": TNTP summary & analysis of USDE draft guidelines*. Retrieved from http://tntp.org/

The New Teacher Project. (2010, February). *Policy Brief.* Retrieved from http://tntp.org/

The New Teacher Project. (2010, November). *How bold is bold?: Responding to Race to the Top with a bold, actionable plan on teacher effectiveness*. Retrieved from http://tntp.org/

The New Teacher Project. (2010). *Teacher Evaluation 2.0*. Retrieved from http://www.tntp.org/files/Teacher-Evaluation-Oct10F.pdf

Tourkin, Warner, Parmer, Cole, Jackson, Zukerberg, et. Al. (2007). *2003-2004 Schools and Staffing Survey (SASS).* National Center for Educational Statistics (NCES) – US Department of Education Institute of Education Sciences.

Tom Torlakson's Task Force on Educator Excellence. (2012). *Greatness by design: Supporting outstanding teaching to sustain a golden state.* A report prepared for the California Department of Education. Retrieved from http://www.cde.ca.gov/eo/in/documents/greatnessfinal.pdf

Trochim, W., & Donnelly, J. (2006). *Research methods knowledge base*. Mason, Ohio: Atomic Dog/Cengage Learning.

Tyack, D. & Cuban, L. (1995). *Tinkering Toward Utopia: A Century of Public School Reform.* Cambridge, MA. Harvard University Press.

UCLA Center X website: www.centerx.gseis.ucla.edu

UTLA Teacher Effectiveness Workgroup. (2012). UTLA's teacher development and evaluation framework. Retrieved from http://turn.learningoptions.net/calturn/archive/utlas-teacher-development.attachment/attachment/TDEF_Framework_march_2012.pdf

U.S. Department of Education's website: http://www.ed.gov/nclb/overview/intro/4pillars.html

U. S. Department of Education, Institute of Education Sciences, "What Works" website: http://www.whatworks.ed.gov

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service. (2006). *Title I Accountability and School Improvement From 2001 to 2004*, Washington, D.C.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development. (2010). *ESEA blueprint for reform*. Washington, D.C.

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, CA: Sage Publications, Inc.

Wei, R.C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the U.S. and abroad*. Dallas: National Staff Development Council. Retrieved from http://www.arts.unco.edu/ciae/institute/2012%20Resouces/2012%20 Jumpdrive%20Resources/Mark%20Hudson/nsdc_profdev_tech_report.pdf

Weisburg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *The New Teacher Project*. Retrieved from http://widgeteffect.org/

Willis, G. (1999). *Cognitive interviewing: A "how to" guide*. Research Triangle Park, NC: Research Triangle Institute.

Yoon, K.S., Duncan, T., Lee, S.W-Y., Scarloss, B., & Shapley, K.L. (2007). Reviewing the evidence on how teacher professional development affects student achievement. *Issues and answers report, REL 2007 – No. 033*. Washington D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://www.ies.ed.gov/ncee/edlabs

York-Barr, J., & Duke, K. (2004). What do we know about teacher leadership? Findings from two decades of scholarship. *Review of Educational Research, 74*(3), 255-316. doi: 10.3102/00346543074003255.