THE RELATIONSHIP BETWEEN TEACHER EVALUATION SCORES AND STUDENT ACHIEVEMENT IN ENGLISH AND MATHEMATICS: EVIDENCE FROM PAKISTAN

by

MUHAMMAD AKRAM

(Under the Direction of Sally J. Zepeda)

ABSTRACT

The purpose of this study was to investigate the relationship between teacher evaluation scores and student achievement in English and mathematics in Pakistan. The goals of the study were to develop a self-assessment instrument to measure teacher evaluation scores, and use those scores to correlate student achievement in English and mathematics in Pakistan. The researcher developed a *Self-assessment Instrument for Teacher Evaluation* (SITE) based on six National Professional Standards for Teachers developed by the Ministry of Education, Pakistan. These Professional Standards were highly compatible with the international research-based teacher quality indicators. Using a convenience sampling method, English or mathematics teachers (*N*=155) of grade 10 in 34 public boys and girls high schools in district Okara were surveyed who self-evaluated their performance on the *Self-assessment Instrument for Teacher Evaluation* (SITE). Additionally, based on the Lahore Board's annual examination results 2012, the student achievement scores in English or mathematics (N=6570) were also collected from these teachers. The data were analyzed using descriptive and inferential statistics. The study found positive, weak or moderate, relationships between teacher evaluation scores and student achievement in English, and essentially no relationship with student achievement in mathematics.

The findings of the study also revealed that Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Effective Communication, and Continuous Professional Development, individually, significantly predicted student achievement in English but not in mathematics. The Subject Matter Knowledge, Instructional Planning and Strategies, and gender significantly combined to predict student achievement in English, explaining 32% of the observed variance in student achievement. Further, the Subject Matter Knowledge and Instructional Planning and Strategies significantly combined to predict student achievement in mathematics, explaining 9% of the observed variance in student achievement. Instructional Planning and Strategies, however, was found to be a mediator, indicating that this variable was uncorrelated with or relatively little related to student achievement in mathematics. Teaching experience did not contribute to student achievement in English and mathematics. The study provided initial evidence of the validity of the SITE.

INDEX WORDS: Teacher evaluation, student achievement, student growth, value-added assessment models, teacher evaluation models and frameworks, self-assessment and teacher evaluation

THE RELATIONSHIP BETWEEN TEACHER EVALUATION SCORES AND STUDENT ACHIEVEMENT IN ENGLISH AND MATHEMATICS: EVIDENCE FROM PAKISTAN

by

MUHAMMAD AKRAM

B.Ed., University of the Punjab, Pakistan, 1995

M.Ed., University of the Punjab, Pakistan, 1998

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

THE RELATIONSHIP BETWEEN TEACHER EVALUATION SCORES AND STUDENT ACHIEVEMENT IN ENGLISH AND MATHEMATICS: EVIDENCE FROM PAKISTAN

by

MUHAMMAD AKRAM

Major Professor: Sally J. Zepeda

Committee: John P. Dayton
Allan S. Cohen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

DEDICATION

I dedicate this dissertation to my parents, Iqbal Begum and Chaudhry Khushi Muhammad (deceased), who taught me the knowledge related to this world as well as the life hereafter, and whose love, care, and prayers have always accompanied me. May Allah (God) shower His blessings on my father and his soul rest in peace (Amen). And my mother! It were you who taught me writing with the *Qalams*—a type of pen made from a dried reed—you used to make for me every day. I pray to Allah Almighty for my parents as He has taught us in the Holy Quran:

> And lower unto them (parents) the wing of submission through mercy, and say: My Lord! Have mercy on them both as they did care for me when I was little (Al-ISRA, 17:24).

This manuscript is also dedicated to my wife (Saima), and children (Zeshan, Rizwan, Imran, and Muneeb) who suffered back in Pakistan during my study but faced the circumstances bravely. I admit that it could not have been possible for me to concentrate on my studies without great cooperation and encouragement of my wife. Saima! I am grateful to you for your sacrifices as well as the efforts you made for upbringing our children. I hope our children will grow inshaAllah (God willing) to be fine men and to offer much to this world. My heartfelt prayers and never ending love will always be with my family.

ACKNOWLEDGEMENTS

This doctoral work was completed with the assistance, encouragement, and great support of many teachers, friends, and relatives. First, I am highly thankful to Dr. Sally Zepeda, major professor, whose continuous, effective, and timely feedback always motivated and pushed me to complete the work well in time. It could not have been possible for me to complete this work without great cooperation of Dr. Sally Zepeda. I have no words to thank her for the long hours she sacrificed for me to complete this work. Dr. Sally! You have been an excellent teacher, a mentor, and a loving and caring friend. May God bless you and shower His favors on you. Amen.

I am also thankful to Dr John Dayton, and Dr Allan Cohen who served in my doctoral committee and assisted me in the best possible ways. Dr Dayton! I am grateful to you for your encouragement and motivation you showed from time to time. Dr Cohen! Your contribution to my study as a methodology professor has been marvelous. Thank you so much both of you.

I cannot forget Dr. Steve Cramer, who assisted me in instrument development process as well as conducting the analysis of the study. He has been continuously assisting me with the research methodology of this study. Thank you so much, Dr. Cramer.

I am thankful to Zafar Masood Anjum who assisted me in data collection from Pakistan. I am also grateful to Abdul Samad, Liaquat Channa, Ghulam Mustafa, Mahmood, and Muhammad Anjum and his family, who stood with me in tough times. Also thankful to my family members and relatives, their prayers and encouragement were a source of motivation. Indeed, you have always been a great source of inspiration.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

The purpose of this study was to investigate the relationship between teacher evaluation scores and student achievement in English and mathematics in Pakistan. Teacher evaluation is a formal and systematic process of examining teacher performance (Stronge, 2006, 2010). The purposes of teacher evaluation are to assess performance of educators not only for certification, tenure and promotion decisions, but also to support valid and legal decisions for termination, and to monitor changes in performance to make improvements where necessary (Darling-Hammond, 1990; Joint Committee on Standards for Educational Evaluation, 1988; Peterson, 2000; Stronge, 2006, 2010; Teddlie, Stringfield, & Burdet, 2003). Comprehensive teacher evaluation systems address accountability and improvement as two wide-ranging purposes which serve the needs of the individual as well as the community. Accountability is central to meet organizational objectives through summative evaluation, while improvement contributes to the professional development needs of individuals through formative evaluation (Barber, 1990; Sanders & Sullins, 2005; Scriven, 1981, 1987; Stronge & Tucker, 1995; Zepeda, 2006, 2012). The combination of both—formative and summative evaluation—results in identifying and supporting effective teachers and also teachers who need targeted improvement.

Regardless of the continent a teacher is geographically located in, evaluating teachers to identify effective and ineffective teachers is a vitally important process to understand. Effective teachers are qualified personnel who demonstrate high levels of teaching expertise, meet the accountability standards, and share professional knowledge with their colleagues (Hunt,

Wiseman, & Touzel, 2009; Loughran, 2006; Stronge & Tucker, 2000). Effective teachers demonstrate competence in subject matter, care deeply about students and their success, and hold distinctive qualities that characterize their effectiveness (Stronge & Tucker, 2000; Wright, Horn, & Sanders, 1997). Effective teachers use their pedagogical skills effectively and enable students to comprehend the content, perform better, and increase their achievement (Brophy & Good, 1986; Wright, Horn, & Sanders, 1997). Stronge and Tucker (2000) argued that perhaps the most significant and empirically tested quality of effective teachers is that they "absolutely, unequivocally, make a difference in student learning" (p. 1).

Researchers believe that using valid measures of student learning in the teacher evaluation process provides accountability evidence, illustrates the influence of the classroom teacher on student learning, and shows substantial effect on student achievement (Mendro, 1998; Sanders & Horn, 1998; Stronge & Tucker, 2000, 2003; Wenglinksy, 2002). This discussion leads to examining the relationship between teacher evaluation scores and student achievement across various subjects at varying levels of schooling in the United States as well as in Pakistan.

In the United States, several researchers and institutions have developed rigorous teacher evaluation frameworks as a basis for developing rubrics for teacher evaluation. The TAP™: System for Teacher and Student Achievement (1999), The Bill and Melinda Gates Foundation's Measures of Effective Teaching (2009), and Robert Marzano's Causal Teacher Evaluation Model (2010) are famous examples of such frameworks. However, Charlotte Danielson's Framework for Teaching (1996), based on its distinctive characteristics, is one of the most widely used frameworks adapted for measuring teacher quality and correlating these measures with student achievement throughout the United States (Gallagher, 2004; Kimball, White, Milanowski, & Borman, 2004; Milanowski, 2004).

Several school systems, Washoe County (Nevada), Cincinnati (Ohio), Vaughn County (Los Angeles), and Coventry (Rhode Island) adapted Danielson's Framework for Teaching (FFT) and developed new teacher evaluation systems in their school districts. A number of quantitative studies have investigated the relationship of teacher evaluation scores on the rubrics used in these counties with value-added scores of student achievement (Kimball et al., 2004; Milanowski, 2004). The results of these studies revealed that most of these teacher evaluation systems had moderate to weak relationships with the value-added scores related to student achievement (Gallagher, 2004; Heneman, Milanowski, Kimball, & Odden, 2006; Kimball et al., 2004; Milanowski, 2004; Odden, 2004). Similar kinds of results were also found from Marzano's studies which showed relatively less significant relationships between teacher evaluation scores and student achievement (Marzano Research Laboratory, 2011).

Given the evidences of weak relationships, did Danielson's FFT-based teacher evaluation systems, the studies based on Marzano's Model, and the studies based on other models adopt comprehensive and effective indicators of teacher quality? Sanders and Rivers (1996) found that effective teachers demonstrated more than a 50 percentile point difference in student achievement as compared to the low performing teachers. If this is true, did the various teacher evaluation systems mentioned develop valid and reliable teacher evaluation instruments that could identify effective teachers? Additionally, were evaluators' ratings valid enough to be used for teacher evaluation? Research indicated that evaluators in the United States are accustomed to rating teachers leniently (Heneman et al., 2006; Kauchak, Peterson, & Driscoll, 1985; Milanowski, 2004; Peterson, 2000), and they may not be able to capture teachers' overall performance on the basis of a limited number of classroom observations that are short in duration (Zepeda, 2012).

One of the most rigorous works on assessing teacher excellence through research-based standards has been summarized by Stronge (2010). He consulted extensive works and studies of various researchers such as Aaronson, Barrow, and Sanders (2007), Danielson (1996), Guskey (2007), Marzano, Pickering, and McTighe (1993), Shulman (1986), and Wenglinksy (2002, 2004), and he summarized eight research-based standards for assessing teacher excellence. These performance standards are aligned with student achievement, contribute to the successful achievement of the goals and objectives, and provide a sound basis for quality of instruction by ensuring accountability for teacher performance (Stronge, 2010). Given the limitations of the evaluators' ratings and based on Stronge's (2010) work, there was a dire need for developing a valid and reliable teacher self-assessment instrument as an alternative that could be associated with student achievement and that could be justifiably attributed to the school or a teacher in the United States (Ballou, Sanders, & Wright, 2004). This study aimed to fill this gap.

Contrary to the research-based evidence in the United States where the emphasis is now transforming from merely evaluating teacher performance to linking teacher performance with student achievement and growth, the teacher evaluation in Pakistan is perhaps the least focused area in its education system (Ministry of Education, 2009). The teachers in Pakistan are evaluated by the school administrators on the Performance Evaluation Report (PER). The PER is a much generalized evaluation report focusing primarily on personality characteristics. Judging the personality characteristics of the individual is important from an ethical perspective but not from a teacher quality point-of-view. The research in the US indicated that "personality characteristics did not necessarily relate to the quality of teaching performance" (Shinkfield & Stufflebeam, 1995, p. 12). This is also true in the Pakistani context as the United Nations Educational Scientific and Cultural Organization (2006) reported that the performance appraisal

system of teachers in Pakistan "is merely a formality… [and it] fails to provide any useful feedback or insight to a teacher's performance" (p. 50).

To meet the challenges faced in the field of teacher education in Pakistan, the Policy and Planning Wing of the Ministry of Education (MoE) implemented a Strengthening Teacher Education in Pakistan (STEP) project in collaboration with the United Nations Educational Scientific and Cultural Organization (UNESCO) in 2008. The STEP project basically focused on developing the Professional Standards for Teachers in Pakistan in consultation with stakeholders in the country. This step was taken as a part of a larger international movement of quality assurance that contributes to the educational quality and impacts student learning outcomes in various fields of human endeavor (Ministry of Education, 2009). These National Professional Standards comprise important teacher evaluation indicators such as Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Classroom Environment, Continuous Professional Development, and others. Since November 2008, these Professional Standards have been formally adopted by all four provincial governments in Pakistan.

The important aspect of these Professional Standards is their relevancy with Danielson's Framework for Teaching (1996), Marzano's causal teacher evaluation model domains (2011), and, especially, the eight standards summarized by Stronge (2010). Stronge has described performance appraisal rubrics for each standard with detailed descriptions from the American perspective. Many of these research-based standards are exactly the same as developed by the Ministry of Education, Pakistan. It was imperative, therefore, to identify which of these Professional Standards were highly effective in the Pakistani public school context. So far, the researcher has not able to find any study which comprehensively addressed the National Professional Standards for Teachers in Pakistan since their development in 2008.

Almani (2002) developed a Teacher Self-Performance Rating Scale (TSPRS) for secondary school teachers—six years before the National Professional Standards were developed by the Ministry of Education Pakistan in 2008—that employed a couple of variables compatible with the National Professional Standards. However, some of the National Professional Standards such as Learning Environment and Continuous Professional Development were not part of the TSPRS. Moreover, the TSPRS was not purposefully developed to correlate teachers' performance with student achievement scores but to measure the effects of in-service training on in-service teachers' performance on certain indicators. In the absence of a valid and reliable teacher effectiveness instrument, it is worthless to discuss the significance and effectiveness of these National Professional Standards. There was an urgent need to develop a new self-assessment instrument for teachers in Pakistan, and this study hopes to help fill this need.

To fill this gap, the researcher developed a valid and reliable instrument—Self-assessment Instrument for Teacher Evaluation (SITE)—for Pakistani public high school teachers, and investigated the relationship between teacher evaluation scores on SITE and 10th graders' achievement in English or mathematics in the 2012 annual examination conducted by the Board of Intermediate and Secondary Education (BISE) Lahore. The researcher hoped this exploratory study would provide not only the base-line data-based evidence of the effectiveness of the National Professional Standards, but also provide the guidelines for further improvement or modification in the Self-assessment Instrument for Teacher Evaluation (SITE). Additionally, the SITE could also be used in American schools as an alternative to evaluators' ratings which have been shown to be lenient, flawed, and biased (Heneman et al., 2006; Kauchak, Peterson, & Driscoll, 1985; Milanowski, 2004; Peterson, 2000).

Statement of the Problem

The literature on the relationship between teacher evaluation scores and student achievement in the United States is plentiful; however, the literature about the relationship between teacher evaluation scores and student achievement in Pakistan is almost non-existent. Research in the US indicates that evaluators are accustomed to rating teachers leniently (Heneman et al., 2006; Milanowski, 2004). Also, evaluators may not be able to capture teachers' overall performance on the basis of a limited number of observations that are short in duration (Zepeda, 2012). Therefore, the teacher evaluation reports based on the evaluators' ratings might be biased, stressful, and disruptive in the United States (Heneman et al., 2006; Kleinman, 1966; Popham, 1971). This is also true in the Pakistani context where teacher evaluations made by school administrators based on the Performance Evaluation Report (PER) are highly problematic as this report is used for seniority purposes rather than on actual performance (UNESCO, 2006).

There is a strong belief that teachers are the best judges of their own performance and can provide data that are not easily captured through any other method (Berk, 2005). Self-assessment is also considered to be the best source for improvement and professional development (Covino & Iwanicki, 1996; Stronge, 2006). Therefore, in the absence of a valid and reliable teacher self-assessment instrument, we might lose important information about teacher effectiveness in the US as well as in Pakistan.

In the Pakistani context, measuring the relationship of teacher evaluation scores with student achievement was required due to various reasons. First, according to the previous result gazettes—books—of the Secondary School Examination (SSE) Lahore, a large number of 10th graders fail every year in various subjects including English and mathematics (Board of Intermediate & Secondary Education Lahore, 2009, 2010, 2011). However, according to the

District Education Officer Lahore, a great majority of the secondary school teachers receive very good evaluations (S. A. Sajid, personal communication, June 15, 2011). In such a situation, the stakeholders such as policymakers, district education authorities, and school administrators cannot link teacher evaluation scores, based on PER, with teacher effectiveness.

Second, the stakeholders cannot determine which factors exhibit more or less levels of teacher effectiveness and the strength of the relationship of those factors with student achievement. Third, researchers, policymakers, district education authorities, and school administrators remain blind of very important teacher quality indicators—professional knowledge, instructional delivery, assessment for learning, the learning environment, professionalism, and teachers' specific role in maximizing student progress—which are absent from the PER. The decisions for teacher promotion, based on PER indicators, therefore, could be highly problematic. And last, but not least, the parents remain deprived of their right that their children must be provided effective teachers.

To point, the National Professional Standards for Teachers in Pakistan, which involve various teacher quality indicators, have not been tested since their adoption in 2008 by the Pakistani Federal government as well as provincial governments. These Professional Standards are high quality indicators and can be the best source of developing a new teacher evaluation system for federal as well as provincial governments. The newly developed Self-assessment Instrument for Teacher Evaluation (SITE) could, perhaps, meet such challenges and provide an alternative approach of measuring teacher effectiveness through the lens of National Standards of teacher quality rather than the PER indicators. The researcher hopes that the SITE would help policymakers and district authorities in making valid decisions about teacher incentives and promotion. Teacher quality in Pakistan is also important to parents so they could make decisions

about placing their children under effective teachers. Lastly, the teachers would be able to assess the weaknesses and strengths in their professionalism and help them focus professional learning on specific areas.

Purpose of the Study

The purpose of this study was to examine the relationship between teacher evaluation scores and 10th graders' achievement in English or mathematics in Pakistan. For teacher evaluation scores, the first type of data, the researcher developed a Self-assessment Instrument for Teacher Evaluation (SITE) that was partly based on the teacher effectiveness indicators and standards as described by Danielson (1996) and Marzano (2010), but was wholly based on the National Standards for Teachers in Pakistan and the standards summarized by James Stronge (2010). The SITE was content-validated, modified, pilot-tested, and used for data collection accordingly. The study sampled those teachers who taught English or mathematics to 10th graders during the academic year 2011-2012 in district Okara, province Punjab. For student achievement scores, the second type of data, the researcher collected 10th graders' achievement scores in English or mathematics obtained in the 2012 annual exams conducted by the Board of Intermediate and Secondary Education (BISE), Lahore. The results were announced in July 2012. The relationships were established on the basis of the teacher evaluation scores and students' achievement scores in English or mathematics.

Background of the Study

Teachers have been evaluated since time immemorial in the United States as well as in Pakistan. In the US, it was, perhaps, 1910, when a teacher was evaluated, for the first time, by a traveling supervisor in Kentucky, and praised with the following words as described by Ellett and Teddlie (2003):

> You were prepared for the lesson, you had different things for young and older students to do, you did not yell and have to spank anyone for being bad, you knew your subjects, the children seemed to get along quite well with you and with each other, you had lots of energy, you did not waste any time telling stories and jokes, and I like you. (p. 104)

Following such a moralistic and an ethical perspective, numerous studies emerged on teacher evaluation between 1920 and 1940. These studies focused on identifying the factors contributing to teaching and the training of prospective teachers (Charters & Waples, 1929).

Shinkfield and Stufflebeam (1995) found that the studies of the National Education Association conducted in 1925 and in Ohio Teaching Records in 1940 introduced numerous teacher evaluation instruments and ratings which, later on, became part of teacher evaluation processes in most of the large cities in the U.S. during the 1960s. During the 1980s, researchers focused on identifying effective teaching methods (Shinkfield & Stufflebeam, 1995) and linking observable teaching practices to a variety of student outcomes (Ellett & Teddlie, 2003).

One of the fundamental concerns with these studies was that they were based on principal's reports, usually recorded on a checklist form, of teacher performance (Peterson, 2000). Although such practices were widely accepted, serious flaws in principals' ratings were identified by researchers and scholars (Kauchak et al., 1985; Medley & Coker, 1987; Peterson, 2000; Stodolsky, 1984; Wise et al., 1984). Medley and Coker (1987) reported low accuracy of principal's judgments and low statistical correlation between administrator's ratings and teachers' roles.

Stodolsky (1984) found that observers' observations were based on a limited number of observations that resulted in unreliable data for evaluation. Others found that principals were not knowledgeable about teacher evaluation and the process of conducting classroom observations; they experienced role conflict in their position of evaluator, and they had inadequate training (Kauchak, et al., 1985; Wise et al., 1984). The research concluded that the relationships of such

ratings with student achievement were also nonsignificant (Bolton, 1973; Castetter, 1971; Coker, Medley, & Soar, 1980; House, 1973; Kleinman, 1966; Popham, 1971). The teachers, therefore, showed severe reactions against appraisal instruments and declared them biased, invalid, and unreliable (Heneman et al., 2006; Kleinman, 1966; Popham, 1971).

In Pakistan, teachers have been evaluated by the school administrators through Annual Confidential Report (ACR) since 1971. With slight customizations, the same ACR had been used in federal and provincial schools for years. In the 1990s, the federal as well as the provincial governments updated Annual Confidential Report (ACR) and implemented it with the name of Performance Evaluation Report (PER). The PER is believed to be fundamentally flawed due to many reasons. First, the purpose of this report was not to provide valid evidence of teacher effectiveness; rather, it aims at providing evidence that a particular teacher possesses "good" personal characteristics and is fit for promotion (UNESCO, 2006). Research in the US illustrates that a teacher's personal attitudes or personality characteristics are not necessarily correlated with his or her teaching effectiveness (Shinkfield & Stufflebeam, 1995). It is implied, therefore, that personal characteristics based teacher evaluation scores on the PER might also not correlate with teacher effectiveness in Pakistan.

Second, the only indicator related to teacher effectiveness recorded in the Performance Evaluation Report (PER) is the pass percentage of students in one subject which is attributed to a particular teacher. In Punjab province, for example, each teacher's pass percentage in one subject is compared with the board's overall pass percentage in the given subject. If the pass percentage of the students in one class in one subject is higher than the overall pass percentage in that subject at the Lahore board's level, for example, the particular teacher of that subject is believed to be highly performing and fit for promotion.

Using the passing percentage as an indicator of teacher quality is problematic as it is based on criterion (33%) which can be achieved with a little extra effort by students. In such case, there are chances that teachers will improve their pass percentage by concentrating on those students who fall just near the pass criterion (33%), and simply ignoring the weaker students. Moreover, based on pass percentage, nobody can identify which student or class demonstrated, for example, a 100% result with a mean of 70, and which student or a class demonstrated a 100% result with a mean of 40. In other words, the mean and the standard deviation in student achievement scores are required to observe the range of the scores and to correlate a teacher's effectiveness score with his or her students' achievement because it gives an idea that a teacher concentrates on every student. The pass percentage, therefore, cannot be realistically used as an alternative to the mean score.

Given the flaws and criticism of the PER, it was important to develop a valid and reliable teacher evaluation instrument for Pakistani teachers. If the Government of Pakistan is committed to improving the quality of education (MoE, 2009) and believes that teachers matter inordinately to student learning (Stronge, 2010), and if the Ministry of Education is committed to "produce world class teachers and empower them to educate the future generations" (MoE, 2009, p. 1), it is essential to provide an in-depth analysis of teacher performance. Through this study, based on the National Professional Standards developed by the Ministry of Education for Pakistani public school teachers, it is hoped that the self-assessment will enhance teacher evaluation. In addition, the researcher also hopes this study might provide to the American education system an alternative source of measuring teacher performance through a self-assessment instrument instead of the supervisor's ratings which have historically been proven to be flawed and biased (Kauchak et al., 1985; Medley & Coker, 1987; Milanowski, 2004; Peterson, 2000).

Research Questions

The following overall research questions were addressed in this study:

1. To what extent do six performance evaluation scales (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development) measured through a self-assessment instrument separately predict student performance in English or mathematics in Pakistan?
2. To what extent do the six scales measured through a self-assessment instrument combine to predict student performance in English or mathematics in Pakistan?
3. Does the addition of teacher gender and teaching experience to the multiple regression model significantly increase the value of prediction in English or mathematics in Pakistan?

Conceptual Framework

This study provides a visual as well as a written product of the interrelated concepts of teacher evaluation scores and their relationship with student achievement in English or mathematics in Pakistan. For measuring the teacher evaluation scores, six National Professional Standards for Pakistani Teachers were selected as a frame of reference. Figure 1 shows the construct of teacher evaluation including the six sub scales (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development). The other type of predictors included teacher's personal characteristics such as teacher gender and teaching experience. Based on the conceptual model, it was assumed that a teacher's score on each domain as well as the six domains (combined) would correlate and predict 10th graders' achievement in English or

mathematics in Pakistan. In addition, based on regression analysis, the researcher assumed that some amount of the observed variance in student achievement in English or mathematics would also be explained by the teacher's gender and teaching experience.

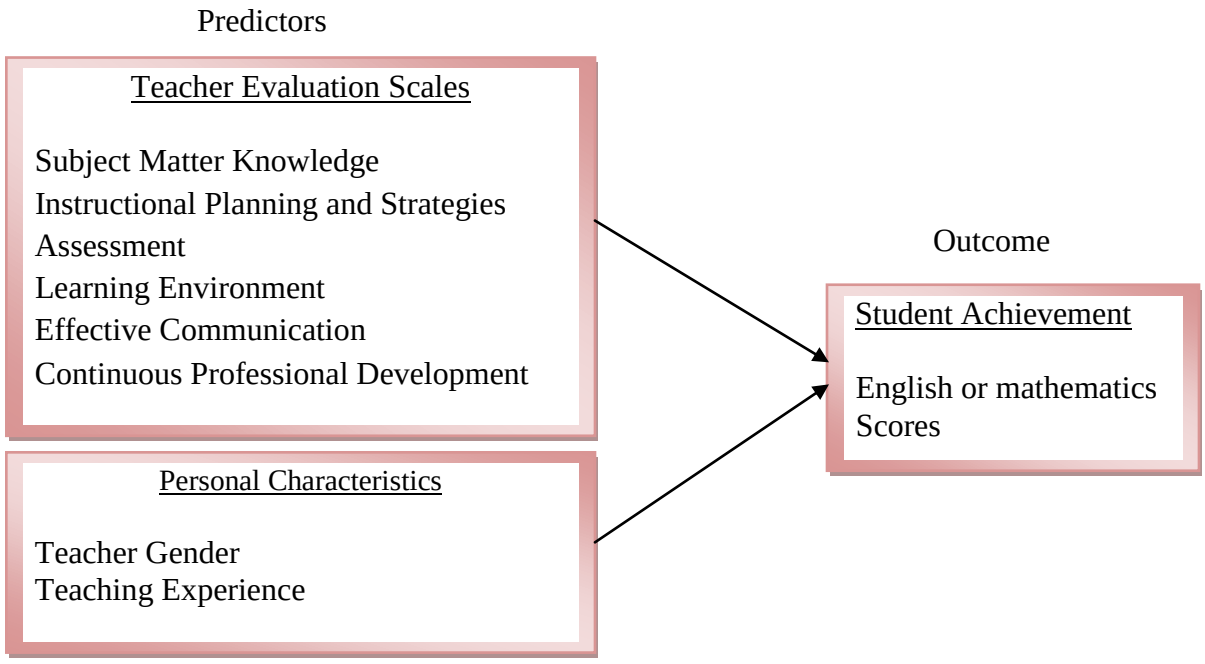

Figure 1

Conceptual Model of the Study

Significance of the Study

A review of the related literature revealed a large amount of research on the relationships between teacher evaluation scores and student achievement in the United States (Borman & Kimball, 2005; Milanowski, 2004). A majority of these studies adapted Danielson's Framework for Teaching (1996) to investigate such relationships which proved to be mixed, relatively small, or not strong (Borman & Kimball, 2005; Kimball et al., 2004; Milanowski, 2004) due to biased ratings of the evaluators (Kleinman, 1966; Popham, 1971) and their inability to evaluate teacher's overall performance on the basis of limited number of observations (Zepeda, 2012). Some of the studies conducted by other researchers, such as Marzano, also provided a small

amount of significant correlations between teacher evaluation indicators and student achievement (Marzano Research laboratory, 2011). Given the biased evidence of evaluators' ratings and their weak correlations with student achievement scores, there was a need for (a) developing a valid and reliable Self-assessment Instrument for Teacher Evaluation (SITE) as an alternative for evaluators' ratings, and (b) using it to measure its relationship with student achievement scores in various subjects. This study addressed this need accordingly.

In Pakistan, various studies pointed out the key issues and problems of teacher education during the last 40 years. There is a general consensus that the quality of teachers is "abysmally low" in the country (MoE, 2009, p. 8). The MoE (2005) Pakistan reported in National Education Census (NEC) that professional preparation of teachers in Pakistan was neither standardized nor based on acceptable professional standards. The MoE (2005) Pakistan also reported lower level of teachers' professional qualification indicating that almost 40% of the teachers had a Bachelor of Education (B.Ed.) Degree, 33% percent held primary teaching certificates in public schools, while almost 40% of the total teaching staff, private schools, in the country were untrained.

Given the problematic situation of poor teacher quality, various institutions such as University of Education, and Institute of Education and Research (IER) have been making efforts to provide professional development trainings to the pre-service as well as in-service teachers across all four provinces (i.e., Punjab, Sindh, Khyber Pakhtunkhwa, and Baluchistan). Additionally, each province has a centralized administrative system of Teacher Professional Development (TPD) which is linked, not closely though, with the district education policymakers and schools (MoE, 2009). For instance, the Directorate of Staff Development established with the name of Education Extension Centre (EEC) in the year 1959, is an apex body of teacher professional development training in the Punjab province which aims at establishing a system of

professional development for teachers and education personnel to enhance the quality of learning in the government schools of Punjab province. The Directorate of Staff Development (DSD) has established a cluster-based program of professional development for the least qualified teachers in the province from 2004 onward.

Like the DSD, other TPD institutions are also focusing on improving teacher quality by providing content skills, collaborative work, mentoring, problem solving, and recognition of the local context in which professional development occurs (UNESCO, 2006). It is significant that these institutions continuously provide professional development opportunities to the teachers so that they can enhance their teaching quality and perform better on the Professional Standards of teacher quality developed by the Ministry of Education, Pakistan. This study hopes to provide a sound basis to those researchers and policymakers who would take interest in measuring teaching quality of those teachers trained by the DSD and other teacher training institutions.

It was not until 2008 that the Ministry of Education (2009) focused on developing a standards-based teacher education and quality assurance system compatible with the international quality assurance indicators which have been shown to have a positive impact on student learning. These National Professional Standards state clearly and succinctly what teachers must know and be able to do, understand what research tells about good teaching and successful learning, reflect the knowledge gained during classroom teaching, and follow professional benchmarks set at acceptable performance levels (Ministry of Education, 2009).

The National Professional Standards demand from teachers an acquisition of current and recent content knowledge of the subjects they teach, the use of broad knowledge of instructional tools and strategies, the ability to monitor and assess student learning outcomes, the ability to communicate effectively with students, parents and other community members, and adhere to a

code of professional conduct (MoE, 2009). These Standards provide a new lens of teacher effectiveness as well as an alternative way of measuring teacher quality where the focus rests upon research-based teacher quality indicators rather than personal characteristics as described in the PER. It was significant, therefore, that these standards be tested through valid and reliable instrument of teacher evaluation. This study hopes to fill this gap by providing initial evidence of teacher quality through a self-assessment instrument based on National Standards.

The current transformation movement toward a standard-based setting was recognized by UNESCO (2006) due to the immediate need for establishing a new teacher evaluation system in Pakistan. In a plea for action related to teacher evaluation in Pakistan, UNESCO suggested:

> A performance based teacher evaluation system and compensation system is required to motivate the teachers to strive toward excellence. Promotion should be linked with teachers' capacities rather than seniority. Additionally there should be an institutional performance appraisal system to monitor institutional accomplishment against set curricular objectives and goals. (p. 55)

The establishment of a performance based new teacher evaluation system required various hierarchical steps to be taken by the provincial governments. First, the teachers must be provided a complete understanding of the professional standards in simple and common language. They must be able to know what they will be expected to do, what will be evaluated, and how the evaluation data would be collected.

Second, there was urgent need for a valid and reliable teacher evaluation instrument that could be formally used in the districts. Since the research in the US suggests that teachers are the best judges of their own performance and principals or evaluators cannot fully capture teacher performance (Zepeda, 2012), there was a need for a teacher self-assessment instrument to be developed in order to explore the worth of these professional standards in the Pakistani context. An immediate beginning toward new teacher evaluation system was required that could help

policymakers to apply the Professional Standards at a large level that could be brought to scale. Interestingly, such an instrument could also be used in the US education system where evaluators' ratings are perceived to be flawed, biased, and lenient (Heneman et al., 2006; Kleinman, 1966; Popham, 1971).

The results of this study could provide several innovative perspectives for policymakers, school administrators, and parents to consider across two continents. In the United States, systems would be able to field test a self-assessment instrument to identify effective teachers, correlate teacher evaluation scores with student achievement and growth, and benefit accordingly. The systems would be able to make valid decisions for teacher's accountability—tenure, removal, incentive, promotion—and professional development would be tailored more on targets with needs (Zepeda, 2012).

In the Pakistani context, this study would provide an informal start toward application of the Professional Standards in the provinces and the districts. The policymakers would be able to know which of these Professional Standards helped identify effective teacher quality indicators. The district and school administrators would be able to identify effective teachers that could help them make accurate decisions about teacher promotion, incentives, and salary. The results of the study could provide initial evidence of the relationship between teacher evaluation scores and student achievement in English or mathematics in Pakistan, and open new opportunities to assess the relationships of teacher quality with student achievement in other subjects such as science, physics, chemistry, and biology. The teachers would be able to identify the current status of their teaching performance and its relation to student learning. Finally, the parents would be able to place their children under effective teachers. Given the various salient features of this newly developed self-assessment instrument, the researcher hopes that this study would provide data-

based evidence of teacher quality and its relationship with student achievement in English or mathematics in Pakistan.

Assumptions of the Study

The study involved the following assumptions:

1. Teachers in public high schools in Pakistan have a basic understanding of the National Professional Standards for Teachers developed by the Ministry of Education.
2. Teachers in high schools will evaluate themselves on the self-assessment instrument fairly and truthfully.
3. All students have equal opportunity to learn English or mathematics in high schools.
4. Students take interest in their learning.
5. Teacher evaluation scores on the self-assessment instrument will predict student achievement in English or mathematics.

Definition of Terms

The study used the following definitions of terms:

Evaluation—The systematic approach to assessing the performance and qualifications related to a particular person in a particular job or professional position (The Joint Committee on Standards for Education, 1988).

Framework for Teaching—A research-based set of components of instruction, aligned to the INTASC standards (Danielson, 1996).

Instruction—Activities or methods teachers use to teach the intended curriculum (Stronge, 2010).

High-stakes testing—Students take tests many times during their educational tenure.

When the score of one test is used to determine sanctions such as retention, placement, or graduation, or rewards such as merit pay, these tests are referred to as high-stakes.

Self-assessment—A concept composed of many elements to produce judgments about one's own teaching for the purpose of self-improvement (Barber, 1990).

Limitations of the Study

The study was limited to the followings:

1. The study sampled only 10th graders and their teachers of English or mathematics in one district (Okara), province Punjab, Pakistan.
2. Instead of using multiple data sources, the study employed Self-assessment Instrument for Teacher Evaluation (SITE) as a single method of data collection.
3. The sample was selected through nonrandomized technique.

Overview of the Research Procedures

It was a correlational study that involved collecting data to determine whether, and to what degree, a relationship existed between two quanfiable variables—teacher evaluation and student achievement. The data were collected from two sources. First, a Self-assessment Instrument for Teacher Evaluation (SITE) was developed, modified, pilot-tested, and used for collecting the evaluation scores of teachers teaching to 10th graders in district Okara, Pakistan. Second, 10th graders' achievement scores in English or mathematics on the Lahore Board's exams conducted in 2012 were also collected from the sampled teachers. Statistical Package for Social Sciences (SPSS) was used for data analyses such as correlation (Pearson r), regression, and t-test for independent samples. The relationships revealed that the scores within a certain range on teacher evaluation were associated with the scores within a certain range on student achievement in English and mathematics.

Organization of the Dissertation

Chapter 1 presents the rationale of this study through the background, statement of the problem, purpose of the study, background of the study, conceptual framework, research questions, and significance of the study. Chapter 2 comprises two parts: first part is related to teacher evaluation in American context that includes historical perspective of teacher evaluation, the literature related to the major areas of teacher evaluation, teacher evaluation models, and frameworks. The chapter also includes literature on value-added models and validity evidences of previous research on teacher evaluation and student achievement. The second part of Chapter 2 discusses teacher evaluation in Pakistan, the Performance Evaluation Report, and the Professional Standards for Pakistani Teachers. In the end, common gaps between American as well as Pakistani teacher evaluation systems are identified.

Chapter 3 describes the methodology of the study, including the framework for the study, the instrumentation and pilot study, population and sampling of the study, data collection, data preparation, data analysis, and limitations of the study. Chapter 4 presents the analyses and findings of the study. Chapter 5 presents an overview of the study, a summary of the findings, principle findings, discussion of the findings, and the implications for policy, practices, and future research.

CHAPTER 2

REVIEW OF THE RELATED LITERATURE

The purpose of this study was to measure the relationship between teacher evaluation scores and 10th graders' achievement in English or mathematics in the annual exam 2012 conducted by the Board of Intermediate and Secondary Education (BISE), in Lahore, Pakistan. The study addressed the following overall research questions:

1. To what extent do six performance evaluation scales (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development) measured through a self-assessment instrument separately predict student performance in English or mathematics in Pakistan?
2. To what extent do the six scales measured through a self-assessment instrument combine to predict student performance in English or mathematics in Pakistan?
3. Does the addition of teacher gender and teaching experience to the multiple regression model significantly increase the value of prediction in English or mathematics in Pakistan?

These research questions were based on The National Professional Standards For Teachers developed by the Ministry of Education, Pakistan to increase the educational quality in the country which has been abysmally low (Hoodbhoy, 1998; UNESCO, 2006). The current Performance Evaluation Report (PER) aims to provide evidence that a particular teacher possesses particular qualities and attitudes and is fit for promotion (UNESCO, 2006). Therefore,

linking the teacher performance scores on the PER indicators with student achievement might be flawed. This study hopes to provide an alternative lens of measuring teacher quality through a valid and reliable self-assessment instrument based on the Professional Standards for teacher quality developed by the Ministry of Education, Pakistan. Additionally, the self-assessment instrument could be used in American schools where evaluators' ratings have been proven flawed (Kauchak et al., 1985; Milanowski, 2004; Peterson, 2000; Wise et al., 1984).

A quantitative approach was selected for this research to describe the extent to which teacher evaluation scores were linked to the students' achievement in English or mathematics in Pakistan. This chapter comprises two parts. The first part discusses key studies that focused on investigating the relationships between teacher evaluation scores and student achievement in the United States; the second part of the review is related to teacher evaluation in Pakistan.

Teacher Evaluation in the United States

Evaluating teaching and identifying effective teachers is not a new phenomenon. Teacher evaluation has a long historical consideration in the United States. It was, perhaps, in 1910 when, for the first time, a traveling supervisor evaluated a teacher in Kentucky and praised in moralistic and ethical ways (Ellett & Teddlie, 2003). Between the 1920 and 1940, numerous studies emerged that focused on the teachers identifying and understanding the factors contributing to both teaching and the training of prospective teachers through ratings (Charters & Waples, 1929). Almost 75 percent of school systems in large cities started using teacher efficiency ratings by 1925 (Shinkfield & Stufflebeam, 1995). The ratings for identifying effective teachers were based on instructional techniques, personality, professional attitude, and maintenance of discipline procedures that incorporated classroom management. Later, through the 20th century,

however, it was realized that "personality characteristics did not necessarily relate to the quality of teacher performance" (Shinkfield & Stufflebeam, 1995, p. 12).

In the 1940s, the faculty members in College of Education at the Ohio State University developed new plans for evaluating classroom teaching through three-stage process: (a) identifying factors of effective teaching, (b) developing instruments for identifying effective teaching, and (c) developing procedure of teacher evaluation which focused on improving teaching rather than merely judging teachers (Raths, 1941). A variety of new instruments had been developed and applied accordingly in the country by 1949.

During 1960s and 1970s, researchers continued focusing on identifying effective teaching methods because of the public demands for quality teaching and increased student learning (Shinkfield & Stufflebeam, 1995). Teachers also showed willingness to be evaluated by their principals (Shinkfield & Stufflebeam, 1995). Teachers' willingness to be evaluated resulted in regular appraisals for their own professional accountability. Researchers also focused on linking observable teaching practices with student outcomes (Ellett & Teddlie, 2003) which urged educational authorities to make teachers accountable for their educational effectiveness.

This increased level of teachers' willingness to be evaluated, however, resulted in using their evaluations for summative purposes such as dismissal, tenure, and promotion (Shinkfield & Stufflebeam, 1995). The school administrators argued in favor of using teacher evaluations for summative purposes as they believed those evaluations directly measured teacher competence (Shinkfield & Stufflebeam, 1995). However, the teachers had doubts about the criteria of judging effective teaching and competence, and had little faith in the validity of teacher appraisal instruments (Bolton, 1973; Popham, 1971). Therefore, teachers demanded for unbiased, objective, credible, and standardized evaluations so that they could be treated fairly and equally.

Teacher evaluation grew considerably in 1980s and evaluation and accountability became the most common buzzwords threaded through educational reforms during this decade (Ellett & Teddlie, 2003). The policymakers focused on evaluating teachers' on-the-job performance for the purpose of licensure at the state level (Ellett & Teddlie, 2003). A new education report, *A Nation at Risk* which emerged in 1983, discussed the issues related to the content, standards, and expectations for students and teachers about their performance. Tyack and Cuban (1995) stated that "the major goal of this legislation was to promote educational excellence and the target was lazy students and incompetent teachers" (p. 78). The remedy was "more discriminating standards for evaluating and compensating teachers, more standardized testing of pupil achievement, more elaborate reporting of test results by local districts to state officials" (Tyack & Cuban, 1995, p. 79). Responding to the *crisis*, the states promulgated more educational laws and regulations than they had generated in the previous 20 years (Tyack & Cuban, 1995).

No Child Left Behind Act is the current federal legislation that emphasized the theories of standards-based education reform (Marsh & Willis, 2007). The standards-based reform was grounded in the belief that by setting high standards and by establishing measurable goals, individual outcomes could be improved. Marsh and Willis (2007) stated that No Child Left Behind (NCLB) actually encouraged "accountability" in public schools, offered parents greater educational options for their children, and helped close the achievement gap between minorities and white students. Marsh and Willis (2007) also stated that NCLB aimed to show achievement toward the goals through federally mandated standardized testing and to measure students' Adequate Yearly Progress (AYP)—an individual state's measure of progress toward the goal of 100 % of students achieving to state academic standards in reading, language arts, and mathematics by 2013-14. NCLB linked state academic standards with student outcomes,

measured student's performances, provided parents with reports about their child's performance, established the foundation for schools and school districts to significantly enhance parental involvement and improved administration through the use of the assessment data to drive decisions on instruction, curriculum and business practices (Marsh & Willis, 2007).

Race to the Top (RTT) is the latest 4.35 billion dollars act created by the United States Department of Education and announced by President Barack Obama on July 24, 2009. Under RTT, the states were awarded points to advance reforms around (a) adopting standards and assessment that prepare students to succeed in college, (b) building data systems that measure student growth, (c) recruiting effective teachers and principals, and (d) turning around the lowest achieving schools (US Department of Education, 2009). The states are, now, receiving millions of dollars to improve teacher quality through developing standards-based systems and bringing instructional improvements into the classrooms.

Responding to the standards-based movements, the states have been focused on linking teacher effectiveness with student achievement, in terms of value-added assessment. Value-added assessment delineates individual student's progress compared against that student's own previous achievement that enables parents and teachers to work together to ensure that students receive the quality education they deserve (Callender, 2004). Value-added assessment involves various advantages as well as shortcomings. It is important, therefore, that the value-added assessment along with its advantages and disadvantages to be understood.

Value-Added Assessment

Value-added, a term originally used in business and economics, has become most widely used as a term to describe certain educational assessment and accountability (Pearson Education, 2008). Value-added assessment refers to any of several models that are being used for

interpreting test scores in a way that evaluates the growth in student achievement over several academic years (Rubin, Stuart, & Zanutto, 2004). William Sanders initially used the term value-added in 1992 for educational assessment and accountability while working on the Tennessee Value-Added Assessment System (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Later on, the value-added assessment drew great attention of researchers and policymakers and became an integral part of various teacher accountability systems throughout the United States.

Value-added assessment is a longitudinal assessment which is different from traditional approaches because it involves complex statistical formulas that are intended to isolate noneducational factors such as the socioeconomic status of students and their demographics to calculate the difference between a student's score for the current year and the previous year. These statistical formulas facilitate administrators and policymakers in identifying effective districts, schools systems, schools, and teachers, accordingly. McCaffrey and Hamilton (2007) described that the school "value-added" values are estimated through complex statistical procedure that uses students' prior test scores to predict their current or future performance. The Tennessee Value-added Assessment Systems, The Dallas Value-added Assessment System, and the Pennsylvania Value-added Assessment system are famous models of this methodology. A comparative analysis of these value-added assessment systems, as described by Stronge and Tucker (2000), Sykes (1997), and McCaffrey and Hamilton (2007), is given in Table 1.

Table 1 shows that the value-added assessment models are statistical models which ask questions such as how much of a change in the student performance can be attributed to students attending one school or one teacher's class. In such models, non-educational factors such as socioeconomic status and demographics are isolated from educational factors using complex statistical formulas. Once these factors are isolated, their impact is removed from measures of

student growth. This growth is deemed true growth and this value-added score is interpreted as a measure of the direct effect of a teacher or a school. The value-added assessment focuses on continuous improvement, fairness, and student improvement.

Table 1

Summary Comparison of the Value-Added Models

	Tennessee Value-added assessment System	Dallas Value-added Assessment System	Pennsylvania Value-added Assessment System
Advantages	• Robust, fair, and a valid measure of student gains • Focus on improvement rather than achievement • Assessment test (TCAP) has good content validity • Positively correlated with evaluations of teachers	• Robust, fair, and a valid measure of student gains • Focus on improvement rather than achievement • The system is good example of technical advances • The goal of the accountability system is student growth	• Robust, fair, and a valid measure of student gains • Focus on improvement rather than achievement • Uses a robust, multivariate, longitudinal mixed effect model in its analyses to yield quality measures of growth
Disadvantages	• Involves complex statistical analyses • Annual testing is a major investment of time, money and human efforts • The TVAAS necessitates only vertically scaled scores of students • There is a potential for misinterpretation, or misuse of data.	• Involves complex statistical analyses • Students are tested multiple times with multiple instruments which can create an emphasis on assessment • There is a potential for misinterpretation or misuse of data.	• Involves complex statistical analyses • Teachers are not engaged with the PVAAS • Lack of effects of the program on student achievement • There is a potential for misinterpretation or misuse of data.

However, researchers now believe that the value-added models fail to address how much a student has grown within a given period of time. Betebenner (2007) stated:

> It is important to note that most value-added analyses focus on the school/teacher level contributions to student growth without first identifying how much a student has grown. Because of this, there is disconnection between these existing growth analysis techniques and determination of how much a student has grown and/or whether a student has made a 'year's growth'. (p. 2)

Further, value-added scores are based on the notion of vertical scaling. Vertical scaling is the process of linking different levels of an assessment which measure the same construct, onto a common scale (Harris, 2007; Holland, 2007).

The value-added measure of student growth might work if the tests measure the same construct over an extended period of time. However, it is reality that the tests are not always vertically scaled where the same constructs are measured in the neighboring grades (Holland, 2007). Instead, numerous new constructs are included in the tests at higher grade levels which are not necessarily aligned to the constructs tested at lower grade level. Therefore, the growth is "best served by considering a normative quantification of student growth" (Betebenner, 2007, p. 4). These drawbacks make it clear that value-added estimates are not the only choice to measure student growth and the search for validity evidence should not be limited to the value-added assessment scores. Rather, student growth percentiles might be used as an alternative to the value-added assessment scores (Betebenner, 2007).

Regardless of the pros and cons of the value-added assessment models, various school districts adopted value-added assessment methodology to measure student growth and link it with teacher evaluation scores. Teacher evaluation scores were collected from some research-based teacher evaluation framework. There is a variety of research-based teacher evaluation frameworks in the US meant for measuring teacher quality. These frameworks have been used

for years and provide some validity evidences of teacher effectiveness. It is important to understand the nature and characteristics of some of the distinguished and research-based teacher evaluation frameworks currently being used.

Teacher Evaluation Models and Frameworks

The key research-based teacher evaluation models/frameworks include:

1. The TAP™: Systems for Teacher and Student Achievement
2. The Bill and Melinda Gates Foundation's Measures of Effective Teaching (MET)
3. Robert Marzano's Causal Teacher Evaluation Model
4. Charlotte Danielson's Framework for Teaching

Each of these four models is briefly discussed with the contemporary research studies that have been conducted by researchers.

TAP™: Systems for Teacher and Student Achievement. The TAP™: System for Teachers and Student Achievement, launched in 1999 by Lowell Milken, focuses on identifying effective teachers through innovative and wide-ranging approaches. Arizona was the first state who implemented TAP in 2000-2001 (Hudson, 2010). The Teacher Advancement Program (TAP) involves four interrelated elements designed around teacher performance, teacher job satisfaction, recruitment, and retention in high schools (Daley & Kim, 2010). The teachers under the Teacher Advancement Programs (TAP) system are provided with (a) Multiple career paths, (b) Ongoing applied professional growth, (c) Instructionally focused accountability, and (d) Performance-based compensation (National Institute for Excellence in Teaching, 2012).

The TAP system partly involves elements taken from Danielson's Framework for Teaching (1996). The TAP system evaluates teachers every year through multiple classroom observations by trained and certified raters and through their contributions to student

achievement growth (Hudson, 2010; National Institute for Excellence in Teaching, 2012). Schacter and Thum (2004) employed an early version of TAP and used observation-based teacher evaluations to correlate with classroom value-added scores. The correlations ranged from .55 (for mathematics) to .70 (for reading). Daley and Kim (2010) also found that:

1. TAP teacher evaluations provide differentiated feedback on teacher performance.
2. TAP classroom evaluations are aligned with value-added student outcomes.
3. TAP teachers become more effective over time.
4. TAP schools show higher retention of more effective teachers, and higher turnover of less effective teachers.

Daley and Kim (2010) implied that teacher evaluation should not be pursued as a one-time event but should be integrated within a comprehensive site-based system, with specific practical elements to support teachers improve teaching and learning in the classroom.

Bill and Melinda Gates Foundation's Measures of Effective Teaching. The Bill and Melinda Gates Foundation, based in Seattle, Washington, is one of the largest private foundations in the world, founded by Bill and his wife, Melinda Gates. It aims to expand educational opportunities and information technology in the United States. In 2009, the Bill and Melinda Gates Foundation launched the Measures of Effective Teaching (MET) project to test new approaches to recognizing effective teaching. The project's goal was to help build fair and reliable systems for teacher observation and feedback to help teachers improve (Bill & Melinda Gates Foundation, 2010). The project included thousands of teachers who helped the foundation to identify better approaches to teacher development through examining multiple measures of teacher effectiveness in various schools. The data were collected from:

1. Students' performance on standardized state and supplemental assessments

2. Video-based classroom observation and teachers' reflections on these lessons
3. Teachers' pedagogical content knowledge
4. Students' perceptions of the instructional environment in the classroom, and
5. Teachers' perceptions of the working conditions and instructional support at their schools (Bill & Melinda Gates Foundation, 2010).

For each type of these data sources, a separate instrument was used. Danielson's Framework for Teaching (FFT) was also used to understand teachers' questioning techniques. The value-added scores of students were collected to measure student achievement gains. The Foundation investigated the relationship between teacher evaluation scores and student value-added scores. Based on the Measures of Effective Teaching (MET) report, a positive relationship, ranging from .38 to .44, was found between teachers' scores on the entire five instruments and 4th-8th graders' value-added scores in the English Language Arts, and mathematics (Bill & Melinda Gates Foundation, 2012). The implications of the study, however, included (a) need for multiple observations when stakes were high, (b) need for multiple measures not just observations or value-added alone, (c) and evidence of large effects of professional development.

Robert Marzano's Causal Teacher Evaluation Model. The Marzano Causal Teacher Evaluation Model is currently being used by the Florida Department of Education as a method that districts can use or adapt as their evaluation model (Marzano Research Laboratory, 2011). Marzano's Causal Teacher Evaluation Model is based on more than 300 experimental and control research studies conducted during the last decade (Marzano, 2003, 2007; Marzano, Frontier, & Livingston, 2011; Marzano, Marzano, & Pickering, 2003; Marzano, Pickering, & Pollock, 2001). The average effect size for classroom strategies was found to be .42 which is associated with a 16 percentile point gain in student achievement. Previously, Marzano's model

had been correlated with student achievement. However, the School District of Indian River County, Florida, planned to employ this model for value-added teacher evaluation in 2011. According to the superintendent of the School District of Indian River County, Florida, Frances J. Adams (2011):

> Beginning in the 2011-2012 school year student assessment results will be incorporated into teacher evaluations. In accordance with F.S.1012.34(3)(a)(1) FCAT scores will be used to measure student growth in learning for classroom teachers whose students take the FCAT for the 2011-2012 school year using Student Growth Approach 1, one of the three models supplied by the state. The Value Added Measure (VAM) for the teacher will be applied. (p. 7)

Based on this statement, Marzano's Model might be used to correlate student growth, in terms of value-added, in the future. According to C. Slezak, the Director of District Partnership of the Learning Sciences International, the Marzano's Model works similar to other value-added models (personal communication, July 20, 2012).

Marzano's Causal Teacher Evaluation Model (2011) is a blend of research and theory correlate with student achievement across four domains:

Domain 1: Classroom strategies and behaviors (41 elements)

Domain 2: Preparing and planning (8 elements)

Domain 3: Reflecting on teaching (5 elements)

Domain 4: Collegiality and professionalism (6 elements)

Marzano's Research Laboratory (MRL) conducted research on the relationship between school improvement and student achievement in Oklahoma in two Phases. During Phase I, research was conducted on nine indicators. Surveys representing indicators for the 9 essential elements were designed, field tested, and sent to students, teachers, and administrators in 61 schools in 2010. The study explored the relationship between teachers' and administrators scores, based on their perceptions, on nine elements and student achievement in mathematics and reading. The results

of the teachers surveys revealed that 5 of the 9 essential elements had significant correlations with the proportion of students proficient or above in mathematics (r=.31 to .39) and reading (r=.33 to .53) (Marzano Research Laboratory, 2010). The results of the administrators' surveys revealed that eight of the nine variables were significantly correlated with student achievement in both subjects. Table 2 shows the summary results of the relationships between teachers' scores and student achievement in English and mathematics.

Table 2

Relationship Between Teacher Scores and Student Achievement in English and Mathematics

	Phase 1	
School Improvement Indicators	English	Mathematics
Curriculum	.39**	.53**
Classroom Evaluation/Assessment	.32*	.41**
Instruction	.30*	29**
School Culture	.31*	.33*
Student, Family, and Community Support	.21	.22
Professional Growth, Development, and Evaluation	.38**	.45**
Leadership	.13	.18
Organizational Structure and Resources	.08	.15
Comprehensive and Effective Planning	.10	.12

*p<0.05, **p<0.01 (Adapted from Marzano Research laboratory, 2010)

During Phase II, 41 instructional strategies were correlated with student achievement scores for reading and mathematics. For reading, 39 of 41 correlations were positive. For mathematics, 41 of 41correlations were positive. However, only 5 out of 41 and 6 out of 41 classroom strategies and behaviors were significantly correlated with reading and mathematics respectively (Marzano Research Laboratory, 2011). When combined 41 strategies into 9 aggregated design questions, the results revealed 6 of those correlations were significant for

reading and only 1 was significant for mathematics. Based on the two reports, it is evident that Marzano's 9 essential elements showed positive association with student achievement in mathematics and reading. Marzano conducted various experimental studies which involved thousands of teachers and students. Marzano's Model is gaining great attention of the researchers and policymakers, and might be used in the future as one of the most powerful tools for linking teacher evaluation scores with student achievement.

Charlotte Danielson's Framework for Teaching. Danielson's Framework for Teaching (1996) is one of the most widely used frameworks for teacher evaluation in the United States. It is a pedagogical model that assists the novice as well as experienced teachers to become effective and efficient teachers. The Framework for Teaching is aligned with the standards set by the Interstate New Teacher Assessment and Support Consortium (INTASC) and compatible with those of the National Board of Professional Teaching Standards (Danielson, 1996). The Framework for Teaching comprises 22 research- based components grouped into four domains of teaching responsibility: (a) Planning and Preparation, (b) Classroom Environment, (c) Instruction, and (d) Professional Responsibilities.

Table 3 shows the 22 components of Danielson's Framework for Teaching. Danielson's (1996) Framework is related to what occurs in the classroom as well as outside of the classroom. Danielson's Framework discusses how to plan for instruction, interact with colleagues, and communicate with parents and the larger community. The Framework is publicly known and "describes those aspects of teaching that occur in some form in every context…although some components are more important in some contexts than in others, the components themselves apply to every setting" (Danielson, 1996, p. 16). Danielson's Framework has been partly used by all the above mentioned teacher evaluation systems, i.e., the TAP Systems for Teacher and

Student Achievement, the Bill & Melinda Gates Foundation's Measures of Effective Teaching (MET), and Marzano's Causal Teacher Evaluation Model.

Table 3

Components of Charlotte Danielson's Framework for Teaching (1996)

Domain 1: Planning and Preparation	**Domain 2: The Classroom Environment**
• Demonstrating knowledge of content and pedagogy • Demonstrating knowledge of students • Selecting instructional outcomes • Demonstrating knowledge of resources • Designing coherent instruction • Designing student assessment	• Creating an environment of respect and rapport • Establishing a culture for learning • Managing classroom procedures • Managing student behavior • Organizing physical space
Domain 3: Instruction	**Domain 4: Professional Responsibilities**
• Communicating with students • Using questioning and discussion techniques • Engaging students in learning • Using assessment in instruction • Demonstrating flexibility and • Responsiveness	• Reflecting on teaching • Maintaining accurate records • Communicating with families • Participating in a professional community • Growing and developing professionally • Demonstrating professionalism

Adapted from Danielson (1996)

There is plenty of research that shows evidence of correlations between the rubrics, based on Danielson's Framework for Teaching, and value-added assessment scores of student outcome in various subjects across varying levels of schooling. The summary results of the studies conducted at school districts of Cincinnati, Coventry, Washoe, and Vaughn are shown in Table 4. According to Table 4, one of the major studies was conducted by Milanowski (2004) who presented the results of an analysis of the relationship between teacher evaluation scores based on multiple data sources and student achievement on district and state tests in reading, and mathematics in Cincinnati.

Table 4

Average Correlations between Teacher Evaluation Scores and Student Achievement

Sites	Grade	Subjects Tested	
Cincinnati		**Reading**	**Mathematics**
2000-01	3-8	.48*	.41*
2002-03	3-8	.28*	.34*
2003-04	3-8	.29*	.22
3 year average:		**.35**	**.33**
Coventry			
1999-00	2-3, 6	.17	.05
2000-01	2-6	.24	-.17
2001-02	4	.39	.34
3 year average:		**.24**	**-.06**
Washoe			
2001-02	3-5	.22*	.20*
2002-03	4-6	.25*	.24*
2003-04	3-6	.19*	.21*
3 year average:		**.22**	**.22**
Vaughn			
2000-01	2-5	.48*	.20
2001-02	2-5	.58*	.42*
2002-03	2-5	.05	.17
3 year average:		**.37**	**.26**

Adapted from Heneman et al. (2006)

Within a value-added framework based on Danielson's Framework for Teaching (1996), Milanowski (2004) correlated the difference between predicted and actual student achievement in reading, mathematics, and science for 3-8 graders with a composite teacher evaluation score based on multiple classroom observations and artifacts collected from the teacher portfolio of 212 teachers. The average intercorrelation between domain scores was .60 and .61 for year 2000-2001 and 2001-2002 with reliability coefficients .86 for both the periods. The study revealed that

teachers' scores from the rigorous teacher evaluation system had moderate positive association with student achievement, from 0 to .5, across all grades and subjects.

White (2004) conducted another study in the Coventry (Rhode Island) School District to investigate the relationship between teacher evaluation scores and student achievement. Within the value-added framework, White (2004) correlated teachers' overall evaluation scores on the rubrics—based on Danielson's Framework for Teaching adopted for low-stakes decisions—with students' achievement on standardized test in reading and mathematics. The researcher sampled 78 teachers of grade 2, 3, 4, and 6 in 2003. The results revealed a small correlation between teacher evaluation scores and reading (.24), and essentially no correlation between teacher evaluation scores and mathematics (.03).

Kimball et al. (2004) described findings from an analysis of the relationship between teachers' scores on a standards-based teacher evaluation system, modeled on Danielson's Framework for Teaching (1996), and student achievement measures in a large Western school district in Washoe County, Nevada. Within the value-added framework, using the evaluators' reports based on the FFT model, Kimball et al. (2004) found initial evidence of a positive but weak association between teacher performance and student achievement in reading, mathematics, and composite scores on standards-based tests.

Gallagher (2004) examined the validity of a performance-based subject specific teacher evaluation system, based on Danielson's Framework for Teaching (1996), by analyzing the relationship between Teacher Evaluation Score (TES) and student achievement in Vaughn County, Los Angeles. Within the value-added framework, Gallagher (2004) investigated whether the Vaughn County teacher evaluation system, designed for high-stakes decisions, had predictive validity of student achievement. The teacher evaluation scores were based on subject-specific

domains with only two domains of the FFT. The student achievement scores were based on the standardized SAT scores. Using the composite score based on multiple data sources, such as observation scores, lesson plans, and documents, Gallagher (2004) found that the teacher's average evaluation score significantly predicted student achievement in Literacy, resulting for each score increase in the TES, with the student achievement showing increases up to 14 percent. However, the relationships between teacher evaluation scores and other subjects such as reading and mathematics were positive but of little practical significance.

Summarizing the results in Table 4, positive but weak relationships between teacher evaluation scores and student achievement scores were found in most of the subjects in various studies. The most and the least stable correlations were found at Washoe and Coventry respectively. The correlations found at Coventry were similar to the results obtained in Washoe County (Kimball et al., 2004), but lower than in Cincinnati (Milanowski, 2004) and Vaughn Charter School (Gallagher, 2004). Based on the weak relationships, Milanowski (2004) and Gallagher (2004) encouraged the researchers to conduct further studies for searching more validity evidence with relatively a larger sample size.

Milanowski's (2004) great interest in exploring the validity evidence of Danielson's Framework for Teaching urged him to explore the construct validity evidence—the missing part in earlier studies—and criterion-related validity of the teacher evaluation ratings based on the four domains of the FFT. For the construct validity, Milanowski's (2004) analyzed the teacher evaluation ratings, as shown in Table 5, in Cincinnati for 2000-2001 and 2001-2002 and demonstrated the correlations among the domain ratings. Most of the correlation coefficients given in Table 5 represent a moderate relationship except between Planning and Professionalism (.75, 77) which are relatively higher. The relationship between Instruction and Classroom

Management is also relatively higher (.68), showing that Professionalism and Planning somehow represent the same content and measure the same construct. The lowest relationship is found between Classroom Management and Professionalism (.43) which shows that these two domains were not similar and measured different variables.

Table 5

Correlation among Domain Ratings, All Cincinnati Teachers Evaluated in 2001-2002 and 2002-2003

Domains	Planning	Classroom Management	Instruction	Professionalism
Planning	----	.56	.56	.77
Classroom Management	.49	----	.68	.54
Instruction	.52	.61	----	.56
Professionalism	.75	.43	.54	----

Upper triangle: Teachers evaluated in 2001-2002 (N=335); Lower triangle, teachers evaluated in 2002-2003 (N=318). Adopted from Milanowski (2004). Used with permission.

Other than domain ratings found at Cincinnati, Milanowski (2004) also presented domain correlations found at Vaughn Elementary School (see Table 6). According to Table 6, high correlations among all domains scores, especially in the subject specific domains (literacy, language development, and mathematics) were found for both years. Milanowski (2004) stated that the subject specific domain ratings have the highest average correlation (.93), the two generic domains have a lower average correlation (.84), and the average intercorrelation of the generic with the subject domains is lowest, but still substantial (.82). The subject specific high correlations demonstrate that subjects are measuring the same construct and providing low construct validity evidence. Milanowski (2004) stated that "these correlations suggest that a

considerable amount of halo effect may be present and suggest the need to examine the rating process and domain rating scales" (p. 7).

Table 6

Correlation among Domain Level Scores, Vaughn Elementary School Teachers Evaluated in 2000-2001 and 2001-2002

Domains	1	2	3	4	5
Lesson Planning	----	.83	.66	.77	.74
Classroom Management	.84	----	.80	.80	.84
Literacy	.84	.92	----	.93	.90
Language Development	.83	.90	.96	----	.90
Mathematics	.84	.86	.95	.96	----

Upper triangle: Teachers evaluated in 2000-01 (N=34); Lower triangle, teachers evaluated in 2001-2002 (N=35). Adopted from Milanowski (2004). Used with permission.

The most important factor of the Danielson's FFT-based studies was related to searching for various types of validity evidence. The researcher carefully reviewed the literature and collected validity evidences from studies conducted at Cincinnati, Coventry, Washoe, and Vaughn. As shown in Table 7, all the sites involved qualified staff to judge the content validity of the rubrics adapted from Danielson's Framework. The studies conducted at Washoe and Coventry did not provide construct validity evidence; the studies conducted at Cincinnati and Vaughn, however, provided relatively higher level of construct-related validity evidence (.60 to .86). Further, all the FFT-based studies provided relatively lower or moderate level of criterion-related validity evidence. The summary results of the validity evidences from all the four sites are shown in Table 7.

Table 7

Validity Evidence of Studies Based on Danielson's Framework for Teaching (1996)

Type of Validity	Cincinnati	Coventry	Washoe	Vaughn
Content Validity	Qualified experts	Qualified experts	Qualified experts	Qualified experts
Construct Validity	Intercorrelation domain average .60	Not Mentioned	Not Mentioned	Intercorrelation domain average .86
Criterion-Related Validity	Moderate relationship, limited Significance	Some positive evidence in reading; not in mathematics	Weak relationship; Significance only in 4 of 9 models	Significant relationship with reading but not with mathematics

Based on FFT-based research studies presented in the review of research, the following potential gaps were identified:

1. Most of the sites customized FFT and used a limited number of its components that limits the results to be generalized over all domains and components of the FFT.
2. The researchers drew relatively smaller sample sizes across each grade. Depending on the teacher qualifications, experiences, and the difficulty levels of the curriculum across each grades, it is imperative that teacher quality must be compared across each grade; a true representative sample of teachers, therefore, must be drawn across each grade.
3. The counties that used teacher evaluation scores for making low-stakes decisions (Washoe and Coventry) demonstrated positive but weak correlation between teacher

evaluation and student achievement. While, the counties that used teacher evaluation scores for high-stakes decisions (Cincinnati and Vaughn) demonstrated relatively higher correlations; these findings demonstrate the possibility of the halo effect pointing possibly to lenient ratings from the principals and assistant principals.

4. Studies conducted at Washoe and Coventry did not provide construct validity evidence.
5. The reasons for weak relationships may include lack of inclusion of comprehensive teacher evaluation indicators, evaluators' role that may be focused on morale improvement rather than performance assessment, and lack of comprehensive teacher performance assessment standards.
6. Supervisor's evaluations are often influenced by a number of non-performance factors such as age, gender of the supervisor and subordinate, and the likability of the subordinate. Moreover, supervisors are not able to capture teacher performance on the basis of observations that are limited in duration.
7. The researchers of the studies conducted at Cincinnati and Washoe feared that teachers were, perhaps, rated leniently which produced higher teacher evaluation scores on the rubrics used for teacher evaluation.

Comparison of the Studies Involved Teacher Evaluation Models and Value-added Models

A comparative analysis of all the four teacher evaluation frameworks and models—The TAP Systems for Teacher and Student Achievement (1999), The Bill and Melinda Gates Foundation's Measures of Effective Teaching (2009), Robert Marzano's Causal Teacher Evaluation Model (2010), and Charlotte Danielson's Framework for Teaching (1996)—was important to understand the overall picture of the teacher evaluation models and their

relationships with students' value-added achievement. Table 8 shows the summary of the comparisons.

Table 8

Comparative Analysis of Teacher Evaluation Systems

Comparison Indicators	**The TAP System**	**Bill & Melinda Gates (MET)**	**Marzano's Causal Model**	**Danielson's FFT**
Involved Danielson's FFT	Yes	Yes	Yes	Danielson's Framework itself
Value-Added	Yes	Yes	Yes	Yes
Focus On	performance-based compensation	high quality videotaped observations	Causal relationship with student outcome	Validity evidences
Sample Size	Large	Large	Small	Small
Nature of Studies	Non-experimental	Non-experimental	Quasi-experimental	Non-experimental
Associated With Student Achievement	Less & weak evidences of association. 2 studies found	Less and weak evidences of association. One study found	Less and weak evidence of association. 3 studies found	Weak evidences of association. More than 4 studies found
Content Validity	Yes	Yes	Yes	Yes
Construct Validity	----------	----------	----------	Relatively higher
Criterion-Related Validity	----------	----------	----------	Relatively moderate
Reliability	High	High	High	High
Teacher Experience Relationship	----------	Not Significant	----------	Not significant

The researcher identified various comparison indicators such as focus of the teacher evaluation models, sample size, nature and strength of association of teacher evaluation systems

with student achievement, and validity evidence across all the four teacher evaluation models that provided a comprehensive overview of the four models. As shown in Table 8, the TAP Systems for Teacher and Student Achievement (1999), The Bill and Melinda Gates Foundation's Measures of Effective Teaching (2009), and Robert Marzano's Causal Teacher Evaluation Model (2010), adapted Charlotte Danielson's' Framework (1996), partly though, and provided little evidence, with only a few studies, of the relationship between teacher evaluation and student achievement. However, Danielson's Framework for Teaching (1996) provided more rigorous and a relatively large number of research-based validity evidence of the relationship between teacher evaluation score and student achievement. Analyzing to all research studies conducted on the relationship between teacher evaluation and student achievement in the United States, the researcher came to the following conclusions:

1. The relationship between teacher evaluation and student achievement in the United States demonstrated weak evidence of validity.
2. Value-added scores were not truly a valid measure of student growth. Student growth percentiles might be used as an alternative of measuring growth (Betebenner, 2007).
3. Gallagher (2004) and Marzano (2010) partly correlated teachers' perceptions and self-assessment scores with student achievement. The results were encouraging. Further studies should be conducted through a valid and reliable teacher self-assessment questionnaire that must include effective teacher equality indicators as identified by Danielson's (1996), Marzano (2010), and Stronge (2010).
4. Teacher experience has not found to be significantly correlated with student achievement. Gender has not been correlated with student achievement in these

studies. It is suggested that the teaching experience and the gender should be correlated with student achievement scores in further studies.

Teacher Evaluation in Pakistan

Pakistan came into existence on August 14, 1947. It comprises four provinces: Punjab, Sindh, Khyber Pakhtunkhwa (KPK), and Baluchistan. Education in Pakistan is essentially a provincial entity and provinces are independent of making provincial level policies and to implement them accordingly. However, to ensure national harmony, to maintain identical education quality standards and indicators, and to preserve the national language and ideological foundations, the Federal Ministry of Education (MoE) is responsible for making country-wide policies. Provinces adopt those policies depending on their contextual needs and situation.

Since 1947, Pakistan has observed more than 15 education policy regimes directing educational improvement in the country (UNESCO, 2006). These policies focused on teacher administration issues and instituting a mechanism for teacher assessment (Kizilbash, 1998), increasing number of teachers, and recruiting teachers and improving the quality of teachers through better pre-service and in-service training (Rahman, Jumani, Akhter, Chisthi, & Jamal, 2011), providing teachers professional development opportunities (Ministry of Education, 1998), and upgrading teacher qualifications (Ministry of Education, 2004). The provinces have been adopting these policies, to a limited extent though, to improve the quality of education. In spite of the federal and the provincial governments' policies, and a growing number of teacher training institutions to support these policies, the quality of teacher and teacher education in the public sector has been abysmally low (UNESCO, 2006).

The issue of low quality of teacher performance was highlighted, for the first time, in the Report of the National Commission on Education 1959 (Kizilbash, 1998). Since then, teacher

quality issues have been consistently addressed by the descendent policies until recently. In a latest analysis of teacher education in Pakistan, the UNESCO (2006) reported that:

> The teacher education programs currently being run by the government institutes are not of the caliber to significantly raise the level of knowledge and skills of teachers to have any measurable impact in the students learning. The curriculum of these programs fails to develop in teachers the required pedagogical skills, subject knowledge, classroom delivery and questioning skills that would make these courses/programs worthwhile. (p. 44)

Such a grim picture of teacher education, according to Aly (2006), Baig (1996), Chaudhry (1990) and various other researchers (as cited in Hoodbhoy, 1998), has been due to various problems such as the political and bureaucratic interferences in education, lack of merit-based appointments of teachers, lack of resources, lack of accountability, lack of internationally comparable learning outcome standards, poorly equipped training institutions, deficient quality of instruction, failure in implementing useful reforms, a defective examination system, and a lack of cost-efficient and high quality teacher and staff training.

Other than these problems associated with teacher education, there is also a perceived consensus that the federal as well as provincial governments have been inconsistent and less concerned in measuring teacher quality. There is only one teacher evaluation report, known as Performance Evaluation Report (PER) that is being used in federal as well as provincial public schools. The Performance Evaluation Report involves various problematic concerns about measuring teacher quality that limit the validity and reliability of this report. A brief description of the Performance Evaluation Report is provided to assist the reader in further understanding these issues in Pakistan.

Performance Evaluation Report

The Performance Evaluation Report (PER) is an official employee evaluation report currently being used in public schools in Pakistan. The basic purpose of the PER is to help

authorities to make decisions about the fitness for promotion of employees (UNESCO, 2006). The Performance Evaluation Report for Secondary School Teachers (SSTs) comprises 8 parts: (I) demographics, (II) personal qualities, (III) attitudes, (IV) proficiency in job, (V) student assessment, (VI) overall grading and fitness for promotion, (VII) remarks of the countersigning officer, and (VIII) adverse remarks by immediate supervisor. Part I is filled by the employee, Part II-VI, and VIII is filled by the school administrator, and Part VII is countersigned by the District Education Officer.

The important portion of the PER is related to the Part II-VI which is filled by the school administrator. Part II and III are related to the personal qualities of teachers such as teacher's intelligence, will power, emotional stability, appearance, and teacher's knowledge of Islam, attitude toward ideology, and relations with superiors, colleagues, and subordinates. Part IV is related to teacher proficiency about their work on various indicators such as power of expression, knowledge of work, analytical ability supervision and guidance, ability to take decisions, and work output and quality. All such indicators are measured through highest to lowest level of scales such as A1, A, B, C, and D. Many of these variables are hard to measure without collecting any kind of data such as teacher intelligence, and knowledge of Islam etc.

A problematic aspect of the PER is that all kinds of variables (from Part II-IV) are based on only school administrator's perceptions and no kind of data such as artifacts are collected from teachers to support administrators' perceptions. The research in the US provides evidence that teachers' personal qualities and attitudes are not necessarily related to the performance of teachers (Shinkfield & Stufflebeam, 1995) and this might be true in the Pakistani context as well. Part V of the Performance Evaluation Report is related to measuring teacher effectiveness through pass percentage of students in one class. If the pass percentage of students in one class in

certain subjects is higher than the Lahore Board's overall pass percentage in that subject, the teacher who taught that subject is deemed as effective or "useful" and the vice versa. The pass percentage, which is based on 33% cut scores, gives no information about the mean and standard deviation of the scores of the students in one class. There is possibility that two classes show identical pass percentage in English, for example, with different mean values such as 40 and 80. In that case, using the mean value for measuring teacher effectiveness would provide totally different results from using the pass percentage for measuring the same construct, i.e., teacher effectiveness. This is, probably, a severe limitation of the Performance Evaluation Report (PER) that hinders researchers from using pass percentages for measuring teacher effectiveness. Summing up, it is evident from the discussion that the PER was developed only for promotion purposes and not for measuring teacher effectiveness.

Contrary to the Performance Evaluation Report, a body of literature, however, exists in the Pakistani context that revealed that the poor quality of teacher education largely impacted teacher quality. One of the earlier studies was conducted by Almani (2002) who compared the effects of in-service training on performance of secondary school teachers in the Hyderabad district in Pakistan. A Teacher's Self Performance Rating Scale (TSPRS) was developed to measure performance of those teachers who had more or less than 10 years of in-service training.

As shown in Table 9, Almani (2002) found that the teacher training significantly affected the classroom performance of female teachers as they performed better in various teacher quality indicators such as teaching methodology, teaching aids, communication style, classroom management, and evaluation. No statistically significant differences were found between male and female teachers in their content knowledge and classroom performance; however, male teachers rated themselves higher than female teachers on motivational techniques. Overall,

teachers with 15 years of experience rated themselves significantly higher than those teachers who had less than 15 years experience on the classroom performance indicators.

Table 9

Research Studies Regarding Teacher Competencies and Teacher Performance in Pakistan

Author	Year	Purpose	Sample	Findings
Almani	2002	To compare effects of in-service training on teacher performance	Secondary School Teachers (N=300)	1. Female teachers rated themselves higher on lesson planning, subject matter knowledge 2. Teachers having more than 15 years experience rated higher on all variables of classroom performance
Bibi	2005	To evaluate the personal as well as professional competencies of secondary school teachers in Pakistan	Heads of teacher training institutions (N=10), teacher trainers (N=50), heads of secondary schools (N=800), and secondary school teachers (4000)	1. Weak competencies in English language 2. Ineffective teaching methods 3. Did not have command over the subject 4. Poor knowledge of the audio visual aids 5. Lack in test construction skills 6. Unable to diagnose the learning difficulties of the students
Jumani	2007	To study teacher competencies	Teachers (N=135), students (N=220), and heads (N=44) for secondary schools, and faculty of education in a university (N=20)	Teachers: 1. were confined to textual knowledge 2. were less competent to present subject matter 3. did not use a variety of teaching strategies 4. lacked in monitoring students' progress 5. did not assess students' work with different techniques

Table 9 (Continued)

Research Studies Regarding Teacher Competencies and Teacher Performance in Pakistan

Author(s)	Year	Purpose	Sample	Findings
Dilshad	2010	To assess quality of teacher education	B.Ed. and M.Ed. students (N=350)	1. Poorly equipped classrooms 2. Lack of highly qualified teachers
Aziz	2010	To find out the effect of demographic factors and teachers' competencies on students' achievement	Heads (N=60), secondary school teachers (N=300), students (N=1500)	Association of student achievement with: 1. teacher planning 2. classroom management 3. teacher experience & evaluating techniques of teachers

The Teacher's Self Performance Rating Scale used in Almani's (2002) study involved various variables of teacher performance such as instructional objectives, teaching aids, child psychology, classroom management, and motivational techniques which were not or less compatible with the National Professional Standards for Pakistani teachers developed in 2008. Some of the variables such as Subject Matter Knowledge or Communication Styles were compatible with the National Standards; however, half of the items under these variables were related to measuring the teachers' perceptions of a good teaching and not their effectiveness. Also, some of the variables given in the National Professional Standards such as Instructional Planning and Strategies, Learning Environment, and Continuous Professional Development were not part of Almani's (2002) scale. Further, some of the items of the TSPRS were grouped into more than one domain that is an important issue of the content validity.

Finally, the purpose of the TSPRS was not to measure the level of frequency of the teacher performance on certain indicators but to measure the level of agreement or disagreement of teachers with their performance based on in-service training. Since the researcher was interested in measuring the level of frequency of teacher performance on certain indicators and not the level of agreement of teachers with their performance, the TSPRS was not a valid measure to use for this study. Due to these various issues and limitations, the author did not use TSPRS for this study.

Bibi (2005) conducted a study to identify and evaluate the personal and professional competencies of secondary school teachers in province Punjab, Pakistan. Ten heads of teacher training institutions, 50 teacher trainers, 800 heads of secondary schools, and 4000 secondary school teachers were randomly selected as a sample. One questionnaire for each type of the sample was developed. The overall results revealed that a significant number of secondary school teachers demonstrated weak competencies in English language, in terms of using grammatically incorrect language while teaching, and used ineffective teaching methods. A majority of the head teachers reported that the secondary school teachers did not have command over the subject they taught, had poor knowledge about the audio visual aids, dealt students in non-psychological ways, did not relate the lessons to daily life experiences, did not have the skills of test construction, and were unable to diagnose the learning difficulties of students.

One of the seminal research studies on teacher competencies in Pakistan was conducted by Jumani (2007). Jumani examined the extent to which teachers trained through distant education possessed competencies in professional knowledge, communication, planning the teaching learning process, assessing student learning, reflecting, evaluating, and planning for continuous improvement. Jumani (2007) found that the aspect of knowledge and understanding

of children were not comprehensively covered in the teacher training programs. Teachers remained stick with the textual knowledge and did not adapt new concepts. Teachers lacked in the ability to structure curricular and co-curricular activities. Jumani's results showed that teachers were neither competent of presenting subject matter, nor did they select appropriate instructional strategies for teaching. Teachers did not use a variety of methods and strategies they learned during training. Jumani (2007) reported that teachers did not provide opportunities to students to apply knowledge, nor did they discuss students' performance issues with students.

Dilshad (2010) conducted a study to assess quality of teacher education in teacher training colleges and various departments of a public university in Pakistan. Dilshad (2010) surveyed 350 student teachers in those colleges and departments of a university and asked about teacher education quality in those colleges and departments on various indicators such as quality of learning environment, quality of contents, and quality of outcomes. Dilshad (2010) found that the low quality of content, lengthy course contents, poorly equipped classrooms, use of English as a medium of instruction, and lack of highly qualified teachers were the main reasons for poor teacher education in the teacher training institutions and departments of the university.

Aziz (2010) analyzed the effects of demographic factors of the students (gender, school context, family size, and income level) and teachers' competencies (Teaching Planning, Teaching Process, Classroom Management, Experience, and Evaluating Techniques) on the achievement of secondary school students. Aziz (2010) sampled 60 head of schools, 300 secondary school teachers, and 1500 students through convenient sampling technique. Three questionnaires were developed, each for heads, teachers, and students. The results revealed that 9th graders' achievement, in terms of pass percentage, was significantly correlated with teacher's scores on planning, classroom management, experience, and evaluating techniques.

One limitation of the Aziz's (2010) study, however, involved using the pass percentage as a measure of student achievement. Usually the pass percentage of one class, which is based on a cut score (33%), is used as students' achievement score which is problematic as it is different from students' actual achievement scores. Take an example of those two students in one class who earn 33% and 80% marks in a certain subject. Since, based on 33% cut score, both students pass, the pass percentage would be 100%. Taking student achievement as 100% in such a case and, then, comparing this achievement with a teacher competency score can be highly problematic as a great deal of information about student scores in terms of standard deviation, and mean score values are missing. The same is true with all students in one class who earn more than 33% marks with varying levels of means scores. To link students' achievement scores, based on pass percentage, with teacher evaluation scores, and making decisions about teacher effectiveness, therefore, might be seriously flawed.

These studies showed that in all of the cases, quantitative approaches were used for data collection and analysis. Findings across studies indicated that most of the teachers were less competent in their content knowledge and monitoring student progress; teachers used poor teaching methodologies, had little knowledge of audio visual aids, and were not able to diagnose the individual needs of their students. Almani (2002) found that teachers having more than 15 years of teaching experience rated themselves higher on all variables of classroom performance, while female teachers rated themselves higher on Lesson Planning, Subject Matter Knowledge, and Child Psychology.

While reviewing the literature for the present study, the researcher found only one study in Pakistani context (Aziz, 2010) that compared scores of teachers' competencies in certain areas (planning, teaching process, classroom management) on the achievement of secondary school

students) with student achievement which was based on overall pass percentage of class. The unit of analysis was a class. In Aziz's (2010) study, the scores of teachers' competencies were positively associated with student pass percentage scores; however, using the pass percentage as a measure of student achievement involves fundamental flaws described above.

As evidenced by these key studies and various reports about poor teacher quality summarized by the UNESCO, the Ministry of Education (MoE) started a review process of the previous education policies and five-year plans in 2005 to launch a new education policy. As the result of that review, the National Education Policy 2009 identified two fundamental gaps—the commitment gap and the implementation gap—as major causes for the weak performance of the education sector during previous years. To address these two gaps, the policy concentrated on widening access to education as well as raising the quality of education as two overarching priorities. In pursuit of these overriding objectives, various policy actions were devised. One of these major action plans was related to improving the quality of education through setting National Standards for educational inputs, processes, and outputs (Ministry of Education, 2009, 2009). The need for setting standards was realized because of the larger international movement of quality assurance in many fields of human endeavor (Ministry of Education, 2009).

The dream of setting National Standards for teachers came true when the Policy and Planning Wing of the Ministry of Education (MoE) in Pakistan in collaboration with the United Nation's Educational Scientific and Cultural Organization implemented Strengthening Teacher Education in Pakistan (STEP) project in 2008 (MoE, 2009). Under the STEP project, National Professional Standards were developed in consultation with stakeholders in all provinces, and adopted by provincial representative in November 2008 (MoE, 2009). The Professional Standards were designed to define competencies and skills deemed to be essential for teachers, to

guide the detailed development of pre and in-service programs of teacher education, and to assure public about the quality of their educators (MoE, 2009). A detailed description of the National Standards for Teachers is important to understand the context of standards and the potential to influence not only increased teacher quality but also teacher evaluation and the potential of teacher self-assessment in both Pakistan and the United Sates.

National Professional Standards for Teachers in Pakistan

The Policy and Planning Wing of the Ministry of Education Pakistan devised the following 10 National Professional Standards for Teachers:

1. Subject Matter Knowledge
2. Human Growth and Development
3. Knowledge of Islamic Ethical Values/Social Life Skills
4. Instructional Planning and Strategies
5. Assessment
6. Learning Environment
7. Effective Communication and Proficient Use of Information Communication Technologies
8. Collaboration and Partnerships
9. Continuous Professional Development and Code of Conduct
10. Teaching of English as Second/Foreign Language (ESL/EFL)

Each of these Professional Standards has 3 parts:

1. Knowledge and understanding of the content
2. Dispositions, and
3. Performance (skills)

The first part—Knowledge and understanding of the content—is related to the content or what teachers know about each of these Standards. Teachers are expected to have deeper knowledge of the standards, new emerging concepts and theories related to each standard, and knowledge of how the learning takes place in the classroom. The second part—dispositions—is related to the teachers behaviors, attitudes, and values they demonstrate against each standard. The third part—performance—is related to the skills of teachers about what teachers can and should be able to do. Combining to three parts, the development of the Professional Standards for teachers is a priority to qualitatively reform the existing system of teacher quality in Pakistan (MoE, 2009).

The important aspect of the National Professional Standards for Pakistani Teachers is their compatibility with the standards being adopted by various school districts in the United States. A majority of these standards is integral part of Danielson's Framework for Teaching (1996), Marzano's (2010) 41 teaching strategies groups into 9 domains, and teacher effectiveness indicators summarized by Stronge (2010). Most of these National Professional Standards are based on the correlational studies, conducted in the US, which provide evidence of positive as well as negative association between teachers' evaluation scores on these standards and students' value-added achievement (Gallagher, 2004; Milanowski, 2004). Some of the most research-based teacher quality standards are being discussed briefly in the following.

Subject Matter Knowledge. One of the fundamental elements of teacher attributes that contribute to student learning and achievement is a teacher's knowledge of the subject matter (Danielson, 1996; Stronge, 2010). The subject matter knowledge refers to the amount and organization of knowledge (Shulman, 1986). Subject Matter Knowledge is a "teacher's understanding of subject facts, concepts, principles, and the methods through which they are integrated cognitively determine the teacher's pedagogical thinking and decision making"

(Stronge, 2010, p. 19). The subject matter knowledge is not only limited to the content knowledge but also it extends beyond to the pedagogical knowledge and curricular knowledge focusing on how to teach and what to teach (Shulman, 1986).

Researchers believe that an effective teacher effectively addresses the appropriate curriculum standards, and integrates key elements and higher-level thinking skills in instruction (Danielson, 1996; Stronge, 2010). Bloom's cognitive taxonomy is a great example of how a teacher can represent a content knowledge. An effective teacher demonstrates accurate knowledge of the subject matter, demonstrates ability to link present content with past and future learning experiences, demonstrates the skills relevant to the subject areas, and understands intellectual, social, emotional, and physical development needs of the age groups (Stronge, 2010). The key elements given above, in combination, provide a picture of effective teaching.

The Ministry of Education (2009) Pakistan necessitates that a teacher must know the basic concepts, theories, and processes of acquiring knowledge of the subject teachers need to teach. Teachers are expected to understand the evolving nature of the subject matter knowledge and keep abreast of new ideas and understanding of teaching and disciplines. Teachers must know the merging concepts, theories, results of researches, and latest trends at national and international levels. The Ministry of Education (2009) also expects that teachers must demonstrate in-depth knowledge of the subject matter and the relationship of that discipline to other content areas as well because linking knowledge of one area to the other is an important factor that deepens a teacher's knowledge. Shulman (1986) stated:

> Teachers must not only be capable of defining for students the accepted truths in a domain. They must also be able to explain why a particular proposition is deemed warranted, why it is worth knowing, and how it relates to other propositions, both within the discipline and without, both in theory and in practice. (p. 9)

This is significant because various disciplines are highly intercorrelated and borrow various concepts from other disciplines. An English teacher, for example, must understand the basic concepts of science because the English curriculum may encompass some or various lessons meant to provide students the information about science vocabulary and knowledge about how day and night comes one after the other, or how the stars revolve around the sun etc. An English teacher with a sound knowledge of the basic scientific concepts, therefore, would be better able to teach those concepts as compared to those teachers who lack or do not have knowledge of such concepts. The knowledge of other disciplines will also be important in subsequent pedagogical judgments regarding relative curricular emphasis (Shulman, 1986).

The research indicates that strong content knowledge of a teacher is positively associated with student learning, especially in mathematics (Aaronson, Barrow, & Sanders, 2007; Goldhaber & Brewer, 2000; Hill, Rowan, & Ball, 2005; Monk & King, 1994; Wenglinsky, 2002). Others found, however, that the subject matter knowledge shows small, statistically insignificant relationships, both positive and negative (Ashton & Crocker, 1987; Haney, Madaus, & Kreitzer, 1987; Quirk, Witten, & Weinberg, 1973). The reason behind the mixed results is that teaching is a multi-faceted activity that encompasses various other factors, especially a teacher's planning strategies, which are equally or more important than merely subject matter knowledge. A deeper understating of planning and teaching strategies is required, therefore, to demonstrate the knowledge accurately.

Instructional Planning and Strategies. Instructional planning and Strategies is another important element of measuring teacher quality and effectiveness. A teacher's teaching begins before a teacher enters into the classroom and starts teaching. Stronge (2010) stated that a teacher's planning of the content, selecting teaching materials, designing the learning activities,

and methods all determine what learning opportunities students are going to have in the classroom. The Ministry of Education Pakistan (2009) keenly addressed this standard while documenting the professional standards for Pakistani teachers and expected that:

> All teachers understand instructional planning, design long-term and short-term plans based upon their knowledge of subject matter, students, community, curriculum goals, and employ a variety of developmentally appropriate strategies in order to promote critical thinking, problem solving, and performance skills of all learners. (p. 12)

Instructional planning and strategies require an effective teacher to use multiple instructional materials, activities, strategies, and assessment techniques to meet students' needs and to maximize their learning (Stronge, 2010; Tomlinson, 1999). Shulman (1986) stated "teachers need knowledge of the strategies most likely to be fruitful in reorganizing the understanding of learners, because those learners are unlikely to appear before them as blank slates" (pp. 9-10). Wenglinsky (2002) stated that effective teachers accept cognitive challenge by providing in-depth explanations of academic content and by covering higher-order concepts and skills thoroughly. Effective teachers also become supportive and persistent in keeping students on task, and they engage, motivate, and maintain students' attention to their lessons (Stronge, 2007).

The research indicates that teachers' instruction and strategies have the most proximal relation with student learning (Cohen, Raudenbush, & Ball, 2003; Marzano, 2007, 2011; Walberg, 1984). Marzano (2011) developed various instructional strategies and conducted over 300 experimental and control studies to investigate the relationship of instructional strategies with student achievement. The average effect size for strategies addressed in the studies was .42. Marzano found that, on average, when teachers used the classroom strategies and behaviors, their typical student achievement increased by 16 percentile points. Various other studies also found similar results (Tomlinson, 1999; Walberg, 1984).

The important aspect of instructional planning and using strategies is related to their appropriate selection according to the need of the learners. An effective teacher knows which of the instructional strategies are more important, depending on the learners' ability to understand those strategies. Many teachers believe homework is the most important strategy to increase student understanding; others believe revising knowledge, asking questions of low-expectancy students, or changing classroom setting are more important factors to enhance students' understanding of the lesson (Marzano, 2011). Regardless of which teaching strategy a teacher uses, the important factor is that the teacher must understands learners' needs and matches to make the strategies compatible with the learners' need, and the relevancy of the strategy with the real-world experiences of the learners (Marzano, 2011).

Assessment. Assessment for learning is a process of evaluating student performance where the teacher gathers, analyzes, and uses data to measure learners' progress (Stronge, 2010). Student assessment provides an overview of what the teacher has taught to the students. Assessment provides diagnostic information regarding students' mental readiness for learning new content, provides formative and summative information needed to monitor student progress, helps keep student motivated, helps students accountable for their own learning, and helps students retain what they have learned (Gronlund, 2006).

The Ministry of Education (2009) requires of teachers that they assess students' learning through using multiple assessment strategies and interpret results to evaluate and promote student achievement. Assessment of student learning can be documented in various ways such as teacher observation, oral questioning, homework assignments, project products, student opinions, criterion-referenced tests or norm-referenced tests (MoE, 2009; Stronge, 2010). The Ministry of Education also requires of teachers that they develop and use teacher made tests for internal

evaluation of students, report assessment data to the parents, and help students engage in objective self-assessment.

Stronge (2010), giving the examples of effective teachers, stated that they use assessment data to develop expectations for students, use a variety of formal and informal assessment strategies, collect and maintain record of student assessment, and develop tools that help students assess their own learning needs. Research indicates that assessment positively influences student learning (Stronge, 2010). Assessment which is aligned with learning targets, accompanied with frequent feedback, involves students deeply in classroom, and is documented properly through record keeping influences student learning (Black & William, 1998; Zacharias, 2007).

Black and William (1998) found that formative assessment has substantial positive effect on student achievement; especially the formative assessment is more effective with low achievers. Guskey (2007) stated that student portfolios were the most important type of assessment tool used to measure student learning. Tomlinson (2007) suggested that teachers must choose the method of assessment that properly fits among students. High stakes testing, however, restricts teachers to formulate approaches of instruction; teachers narrow the curriculum, and focus on memorization, drills and worksheets, and allocate less time to higher order thinking skills (Stronge & Xu, 2011).

Learning Environment. Students need an engaging and stimulation learning environment to support student proper growth (Stronge, 2010). Effective teachers create an environment of respect and rapport in their classrooms by the ways they interact with students and by the interaction they encourage and cultivate among students (Danielson, 1996). Effective teachers focus on the organization of learning activities throughout teaching and learning, maximize instructional time, assume responsibility for student learning, and establish rapport and

trustworthiness with students by being fair, caring, and respectful (Good & Brophy, 1997; Marzano, Marzano, & Pickering, 2003; Wang, Haertel, & Walberg, 1994).

The Ministry of Education (2009) described that effective teachers create a supportive, safe, and respectful learning environment that encourages positive social interaction, and active engagement in leaning and self-motivation. As a result of this interaction, a positive learning environment can shape student outcomes in cognitive, motivational, emotional, and behavioral domains (Ludtke, Robitzsch, Trautwein, & Kunter, 2009). Danielson (1996) stated that classrooms with a positive climate for learning are cognitively busy places, with students and teacher setting a high value on high-quality work.

Research indicates that in a positive learning environment, effective teachers develop functional floor plans and material placement for optimal benefit, and establish classroom rules and procedures (Emmer, Everston, & Worsham, 2003; Stronge, 2007). Kunter, Baumert, and Koller (2007) found that the students' perceptions of rule clarity and teacher monitoring are positively related to their development of academic interest in secondary school mathematics classes. Effective teachers have less disruptive student behaviors than do less effective teachers (Stronge, Ward, Tucker, & Hindman, 2008). Wang, Haertel, and Walberg (1994) found that classroom instruction and climate was the second most influential factor among six identified types of influence, second to student aptitude. Summarizing to these findings, a positive classroom environment increases student-teacher interaction, maximizes instructional time, and helps students improve their achievement.

Effective Communication. The ability to communicate is one of the essential requisites for teacher effectiveness (Fullan, 1993). Communication is an ability to (a) package and deliver content meaningfully, (b) create an engaging class culture, (c) be sensitive to individual student

needs, and (d) connect with the student, first, as a person, and, then, as a learner (Cornett-DeVito & Worley, 2005). Stronge and Tucker (2003) stated that effective teachers communicate effectively with students, model standard language (English), actively listen and respond in a constructive manner, establish and maintain multiple modes of communication between school and home, and adhere to school district policies regarding communication of student information. Effective teachers use knowledge of effective verbal, nonverbal, and written communication techniques and tools, and collaborate and support interactions with students and parents (MoE, 2009). Effective teachers also explain concepts in simple and logical sequence, and explain lessons according to the age and ability of the students (Stronge, 2010).

Research indicated that students taught by teachers with greater verbal ability learn more than those taught by teachers with lower verbal ability (Cornett-DeVito & Worley, 2005). Catt, Miller, and Schallenkamp (2007) interviewed 11 award winning teachers to develop a better understanding of their instructional communication practices. The authors found that those teachers understood the ebb-and-flow of the classroom and allowed spontaneity, used a wide range of communication skills, and created relationships with students to establish interpersonal rapport (Stronge, 2010). Catt et al. (2007) also encouraged an open, warm, and communicated environment that invited students' comments. The results of the Catt et al. (2007) study revealed that open and warm communication with the students, parents, and community helped teachers as well as students perform better. These findings show that effective teachers can maximize student learning though discussing students' problems with their colleagues, and adapt those behaviors followed by the teachers better in communicating with students.

Continuous Professional Development. Professional development is a process of improvement in which teachers participate as active and responsible members of the professional

community, engage in reflective practices, pursue opportunities to grow professionally, and establish collegial relationship to enhance the teaching and learning process (MoE, 2009). Effective teachers value and practice the principles, standards, ethics, and legal responsibilities of teaching, and monitor and strengthen the connection between their own development and student development (Fullan, 1993). Effective teachers maintain a professional demeanor and appearance; they adhere to professional standards, use self-assessment strategies to improve performance, and explore knowledge about effective methods (Stronge & Tucker, 2003).

The research indicated that good teachers care about their students; resultantly, students respond to the teachers by optimizing their commitment to learning (Lumpkin, 2007). The results of a meta-analysis revealed that teachers who continue to receive professional development can boost their students' achievement up to 21 percentile points (Yoon, Duncan, Lee, Scarloss, & Shapley (2007). Guskey (2002) found that professional behaviors of effective teachers encourage linking professional growth goals to professional development opportunities. These findings show that effective teachers act individually and collectively to advance the teaching profession, act as shapers, and well informed critics of educational policies, instructional innovations, and internal changes that impact student learning (Little, 1993; Stronge, 2010).

The literature related to the National Professional Standards and their relationship with student achievement requires that these standards must be tested in the public schools in Pakistan. The researcher has not been able, so far, to find any research study, including Almani's (2002) that used questionnaire methods that encompassed the National Standards for Pakistani teachers and measured teacher quality. This study was designed to fill this gap. The researcher developed a Self-assessment Instrument for Teacher Evaluation (SITE) for Pakistani public high

school teachers and used it to correlate 10th graders' achievement in English or mathematics on the Lahore Board's annual examination 2012, in one district, Okara, province Punjab.

Examining the Teacher Evaluation in the United States and Pakistan

Based on the literature on teacher evaluation and student achievement in the United States and in Pakistan, the following two gaps emerged:

1. From United States' perspective, the relationships between teacher evaluation scores, based on principal/assistant principal or observer's ratings, and student value-added assessment scores were mixed, relatively small, or not strong (Borman & Kimball, 2005; Gallagher, 2004; Kimball et al., 2004; Milanowski, 2004). Many teachers raised voices against evaluator's competence, strictness, and leniency (Heneman et al., 2006; Milanowski, 2004).
2. From Pakistani perspective, based on headmasters' and students' reports about teachers' competencies, teachers were less competent in various teacher quality indicators. Teachers had little knowledge of the content and audio visual aids, lacked in test construction skills, and had little knowledge of different teaching methodologies (Aziz, 2010; Bibi, 2005; Jumani, 2007).

To deal with these two types of gaps, the researcher argued in favor of the following:

1. A self-assessment instrument, based on the standards common among the Danielson Framework for Teaching (1996), Marzano's Model, Stronge's work (2010), and the National Standards for Pakistani Teachers, must be developed as an alternative to the ratings of principals/assistant principals/headmasters/headmistresses.
2. The self-assessment instrument must be used for collecting teacher evaluation scores.

3. The teachers' evaluation scores on the self-assessment instrument must be correlated with student's scores calculated through Student Growth Percentiles in the US.
4. The teachers' evaluation scores on the self-assessment instrument must be correlated with students' achievement scores in certain subjects in Pakistan.

Teacher Self-Assessment

Why the researcher argued in favor of using a self-assessment instrument for teacher evaluation is based on the literature that supports the idea of using a self-assessment tool as an opportunity for one's self-improvement and professional development (Centra, 1973, 1977; Peterson, 2000). Self-assessment is a very powerful tool for measuring teacher quality as side by side using the ratings done by principals or other administrators (Danielson, 1996, Peterson, 2000). Principals or administrators judge teachers' performance through observation and complete ratings or checklists during observation process (Darling-Hammond, Wise, & Pease, 1983, Medley & Coker, 1987). Rating the teachers on the basis of limited observations and then generalizing those ratings over their overall teaching performance provides limited evidence of reliability (Zepeda, 2012). It is quite possible that during those observations teachers were well prepared and demonstrated high performance, or they were stuck with some serious social problems and demonstrated very low or average performance. Supervisors, therefore, can only capture limited sample of teachers' teaching performance through observation (Zepeda, 2012).

Studies show that supervisor evaluations are often influenced by a number of non-performance factors such as the age and gender of the supervisor and subordinate and the likability of the subordinate (Alexander & Wilkins, 1982; Bolino & Turnley, 2003; Heneman, Greenberger, & Anonyuo, 1989; Varma & Stroh, 2001). Moreover, principals are generally effective at identifying the best and the worst teachers but not able to distinguish teachers in the

middle of the achievement distribution (Jacob & Lefgren, 2005). Further, supervisors are vulnerable to teachers' reactions in terms of subject matter expertise, school context, peer evaluation, use of portfolio, evaluator's competency, strictness, and leniency in ratings (Heneman et al., 2006; Milanowski & Heneman, 2001).

Teacher self-evaluation, on the other hand, is a frequently advocated data source for teacher evaluation (Barber, 1990; Bodine, 1973; Carroll, 1981; McGreal, 1983; Peterson, 2000). The self-assessment is a process in which teachers make judgments about the adequacy and effectiveness of their own knowledge, performance, and pedagogical skills for the purpose of self-improvement (Airasian & Gullickson, 1997). Research indicated that teachers do monitor and improve their own behavior in relation to goals, expectation, and outcomes, act on self-gained data, and engage themselves in professional development activities (Barber, 1990; Festinger, 1954; Peterson, 2000).

Teacher self-assessment makes teachers aware of their strengths and weaknesses, encourages collegial interactions and teacher development, assists in school improvement, and helps administrators in making decisions about teaching assignments (Peterson, 2000). Self-assessment gives teachers' control over their own growth and treats teachers as professionals (Airasian & Gullickson, 1997). As demonstrated by some of the studies, teachers, by themselves, are the best judges of their teaching performance and growth (Airasian & Gullickson, 2006; Clandinin & Connelly, 1988; Stufflebeam & Shinkfield, 1985).

Danielson (1996) recommended that "the most powerful use of the framework (for teaching), and one which should accompany any other use, is for reflection and self-assessment" (p. 53). Though there is possibility that experienced teachers would rate themselves higher on teaching effectiveness indicators as Almani (2002) found in his study, a self-assessment evidence

can provide support for what teachers do in the classroom and can present a picture of their teaching unobtainable from any other sources (Berk, 2005). Also, teachers are more likely to act on self-gained data than on information from other resources (Centra, 1972). Moreover, teachers' perceptions would be based on multiple data sources such as samples of students' work, logs of professional development activities, and contacts with families which are important elements of the teacher quality indicators. Lastly, collecting data through teachers' self-assessments is feasible, cost efficient, and time saving (Goe, Bell, & Little, 2008). The researcher, therefore, developed and then used the Self-Assessment Instrument for Teacher Evaluation (SITE) as a single method of data collection for this study. The researcher hopes that the self-assessment instrument might serve as an alternative to the ratings of principals and school administrators.

CHAPTER 3

METHODOLOGY

The purpose of this study was to measure the relationship between teacher evaluation scores and 10th graders' achievement in English or mathematics in the 2012 annual exam conducted by the Board of Intermediate and Secondary Education (BISE) Lahore, Pakistan. The study addressed the following overall research questions:

1. To what extent do six performance evaluation scales (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development) measured through a self-assessment instrument separately predict student performance in English or mathematics in Pakistan?
2. To what extent do the six scales measured through a self-assessment instrument combine to predict student performance in English or mathematics in Pakistan?
3. Does the addition of teacher gender and teaching experience to the multiple regression model significantly increase the value of prediction in English or mathematics in Pakistan?

This chapter is divided into eight sections to describe the methodology employed in answering the questions directed to investigate the relationship between teacher evaluation and student achievement. The first section is related to the conceptual framework of the study. Section two describes the construction of the Self-assessment Instrument for Teacher Evaluation (SITE) including the description of the pilot study. Section three describes selection of the

sample from the target population. Section four describes the student achievement scores. Section five describes procedure of data collection. Section six is about data preparation. Section seven discusses the data analysis, and section eight is about the limitations of the study. It is vitally important to understand how the data collection procedure was adopted and reported.

Conceptual Framework

The purpose of this study was to investigate the relationship of teacher evaluation scores with student achievement. Teacher evaluation is a formal and systematic process of examining teacher performance (Stronge, 2006). One of the purposes of teacher evaluation is to identify high quality teachers (Peterson, 2000; Stronge & Tucker, 2003). Identification of high quality teachers is important because effective teachers are believed to use their pedagogical skills effectively, enable students to comprehend the content, perform better, and improve student achievement (Brophy & Good, 1986; Sanders & Rivers, 1996; Wright, Horn, & Sanders, 1997).

There is a considerable theory that specifies how a variety of indicators of teacher quality can be grouped into various teacher evaluation models. Teacher evaluation is a complex phenomenon that involves multifaceted procedures, aspects, and contexts. Therefore, it is hard to measure this phenomenon comprehensively through a single teacher evaluation model or theory. Since this study was not designed to support or oppose any particular theory, the researcher created his own theoretical framework that supported this study.

Through a careful review of various teacher evaluation models such as (1) Charlotte Danielson's Framework for Teaching (1996), (2) Robert Marzano's Causal Teacher Evaluation Model (2010), (3) James Stronge's (2010) work on teacher effectiveness standards, and (4) the National Professional Standards for Teachers in Pakistan (2009), the researcher initially clarified the construct that would guide the self-assessment instrument development. Through a

comparative analysis of the four types of teacher effectiveness models and works, the researcher particularly selected those standards which were necessarily part of the National Professional Standards For Pakistani Teachers because of the context of the study, as well as they were compatible with the research-based standards employed by the various standards-based teacher evaluation models in the US, especially as summarized by Stronge (2010). After a careful comparison, the following six most common components of the construct of teacher evaluation were finally selected for the instrument development:

1. Subject Matter Knowledge
2. Instructional Planning and Strategies
3. Assessment
4. Learning Environment
5. Effective Communication
6. Continuous Professional Development

The six domains served as the major variables of the study. The study endeavored to investigate the extent to which teacher evaluation scores measured through the six main variables would predict student achievement in English or mathematics as well as the additional increment of prediction in English or mathematics through personal characteristics of teachers such as teacher gender and teacher experience. Therefore, two additional variables, teacher gender and teaching experience, were also included in the conceptual model (see Figure 2).

Conceptual Model

The relationships of teacher evaluation scores, teacher gender, and teaching experience with student achievement scores in English and mathematics are shown in Figure 2. According to the figure, two types of variables, teacher evaluation variables (Subject Matter Knowledge,

Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development), and personal characteristics (teacher gender and teaching experience) were selected as predictors. Student achievement in English and mathematics were selected as outcome variables. The purpose of the first and second research question was to investigate the extent to which the six subscales of teacher evaluation scores would predict students' achievement, separately and combined, in English or mathematics. The third research question was designed to investigate additional amount of variance in student achievement scores in English or mathematics explained by teacher gender and experience.

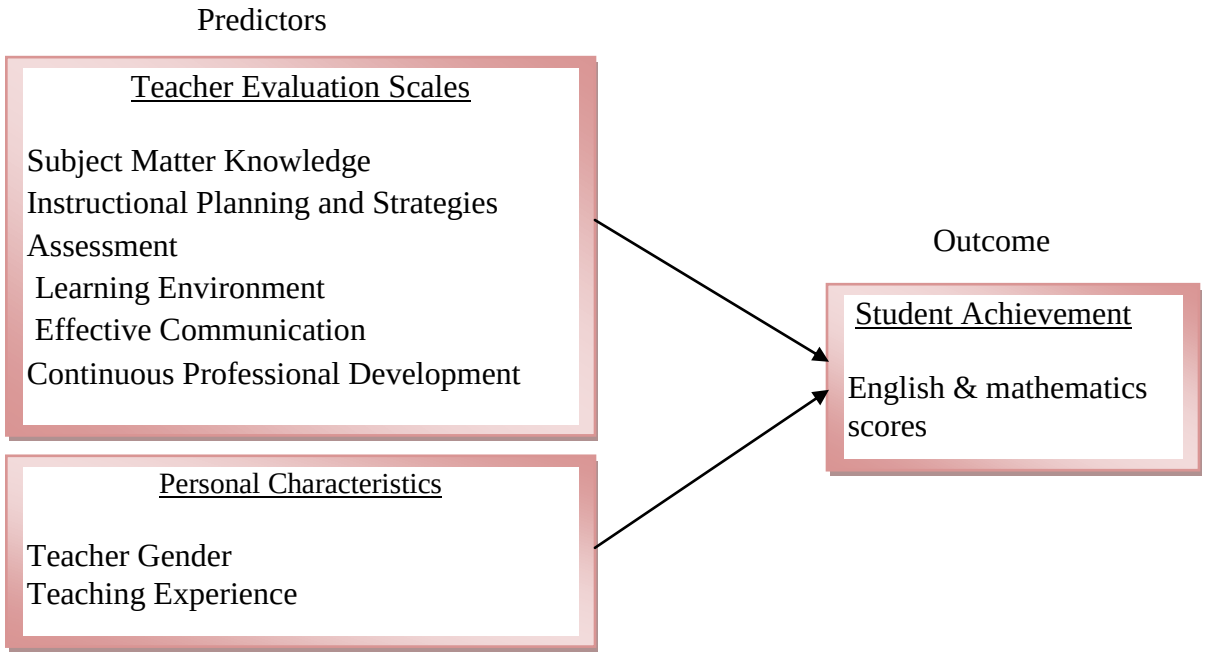

Figure 2

Conceptual Model of the Study

Descriptive statistics, Pearson correlations, simple linear regression, multiple regression, and t-test for Independent Samples were conducted to explain the relationships of teacher evaluation domains with student achievement in English or mathematics using the SPSS 17th version.

Instrumentation

Measuring the teacher effectiveness is a complex phenomenon as it involves various contextual issues that support the idea that one size does not fit all. The researcher thoroughly delved into the teacher evaluation literature in Pakistan to find a valid and a reliable teacher evaluation instrument based on the National Professional Standards for Pakistani Teachers that could be adapted for this study. Similarly, the researcher rigorously searched literature on teacher evaluation in American perspective and found various rubrics based on teacher evaluation models such as Danielson's framework for Teaching (1996), Marzano's Causal Teacher Evaluation Model (2010), and Bill and Melinda Gates Measures of Effective Teaching (2009) adapted by various local school districts according to their contexts and needs. The researcher was able to find one self-assessment questionnaire—Teacher Self-Performance Rating Scale (TSPRS)—developed by Almani (2002) in Pakistan that demonstrated little relevance with the teacher quality indicators compatible with the National Standards for Pakistani Teachers.

The Teacher Self-Performance Rating Scale (TSPRS) was developed six years before the Ministry of Education (MoE) developed National Professional Standards in 2008. Half of the items in the TSPRS were not developed to measure the teacher effectiveness but to measure teachers' perceptions of a good teaching, while other important variables which were integral part of the National Professional Standards such as Learning Environment and Continuous Professional Development were not part of the TSPRS. Since the TSPRS did not comprehensively cover the National Professional Standards, the researcher used James Stronge's (2010) research-based standards for assessing teacher excellence as an alternative and developed a Self-assessment Instrument for Teacher Evaluation (SITE) with the help of teacher evaluation content specialists and practitioners. The SITE comprehensively covered a major portion of the

National Professional Standards for Pakistani Teachers. The final version of the SITE is in Appendix A.

The largest single contribution to the SITE came from extensive work of Stronge (2010). Stronge provided a detailed description of each of the given standards, the key areas of standards, sample definitions of the standards, and what was supported by the seminal research works for that particular standard. Table 10 provides summary of the research-based evidences found for each standard.

Table 10

Key References for Six Teacher Evaluation Standards

Standards	Research-Based Evidences
Subject Matter Knowledge	Aaronson, Barrow, & Sanders (2007); Cornett-DeVito & Worley (2005); Goldhaber & Brewer (2000); Hill, Rowan, & Ball (2005); Monk & King (1994); Wenglinsky (2002).
Instructional Planning and Strategies	Buttram & Waters (1997); Covino & Iwanicki (1996); Johnson (1997); Stronge (2007); Tomlinson (1999); Walberg (1984); Wenglinsky (2002).
Assessment	Black & William (1998); Gronlund (2006); Guskey (2007); Tomlinson (2007); Zacharias (2007).
Learning Environment	Emmer, Everston, & Worsham (2003); Good & Brophy (1997); Ludtke, Robitzsch, Trautwein, & Kunter (2009); Wang, Haertel, & Walberg (1994).
Effective Communication	Catt et al. (2007); Cornett-DeVito & Worley (2005); Fullan (1993); Covino & Iwanicki (1996); Sachs (2001).
Continuous Professional Development	Fullan (1993); Guskey (2002); Little (1993); Lumpkin (2007); Stronge & Tucker (2003); Yoon, Duncan, Lee, Scarloss, & Shapley (2007).

Besides the research-based evidences, Stronge (2010) used teacher quality indicators which are "tangible behaviors that can be observed or documented to determine the degree to which a teacher is fulfilling" the particular standard (p. 23). The researcher adapted those tangible behaviors as teacher quality indicators and used them for instrument development. The instrument development process is given in Table 11.

Table 11

Self-Assessment Instrument for Teacher Evaluation (SITE) Development Process

Concept Clarification
Item Pool Development and Refinement
Response Scale Development
Open-ended Question
Selecting Predictor Variables
Expert Review and Final Refinement of Survey Instrument
Pilot Study

To understand the methodology of the study, each of these processes is important to understand.

Concept Clarification

Concept clarification is an important process of instrumentation. The basic purpose of the concept clarification is to provide a deeper understanding of each theoretical concept or construct to the reader. The labels *concept* and *construct* will be used interchangeably. Waltz, Strickland, and Lenz (1984) suggested that the investigators must translate an informal working definition into a theoretical definition that is precise, understandable to others, and appropriate in the context in which the term will be used. The concepts can be defined through analysis, synthesis,

and derivation (Walker & Avant, 1983). A concept analysis is used when a body of literature on the concept is available. A concept synthesis is done based on clinical observation, and, a concept derivation is used when a concept moves from one field of interest to another. Since a plethora of research on concept definitions was available, the researcher used the *concept analysis* method for concept definitions. The researcher delved into the literature and found sample definitions from Stronge's (2010) work that could be used to help operationalize the teacher performance standards. The concepts along with their definitions are shown in Table 12.

Table 12

Definitions of Teacher Evaluation Components

Theoretical Components	Definitions
Subject Matter Knowledge	A teacher's understanding of subject facts, concepts, principles, and the methods through which they are integrated cognitively (p. 19).
Instructional Planning and Strategies	The teacher uses appropriate curricula, instructional strategies, and resources during the planning process to address the diverse needs of students (p. 33).
Assessment	The teacher gathers, analyzes, and uses data…to measure learner progress, guide instruction, and provide timely feedback (p. 56).
Learning Environment	The teacher creates and maintains a safe classroom environment while encouraging fairness, respect, and enthusiasm (p. 66).
Effective Communication	The teacher communicates effectively with students, school personnel, families, and the community (p. 76).
Continuous Professional Development	The teacher maintains a professional demeanor, participates in professional growth opportunities, and contributes to the profession (p. 86).

Adapted from Stronge (2010)

Item Pool Development and Refinement

Concepts vary in the extent to which the domain of the observable indicators is either large or small (Brink & Wood, 1998). Item development, therefore, depends on the breadth of the domain of the observable indicators. Given the six domains selected for this study, the researcher focused on adapting those observable indicators—teacher quality indicators—found in Stronge's (2010) work, which are relatively very broad. These sample quality indicators are highly compatible with the definitions of the standards given in concept clarification process. Going through each quality indicator, the researcher carefully selected 44 items grouped into six domains and used them to develop response scale.

Response Scale Development

Constructs can be measured through various types of response scales. Some of the response scales measure constructs dichotomously, others polytomously (Yen, & Fitzpatrick, 2006) and use Likert or Likert-like scales that measure level of agreement, level of acceptability, level of desirability, or level of priority and various other response scales depending on the purpose of the research. For this study, the researcher used polytomous scales to measure the levels of frequency, from lowest to highest, against the items grouped into six subscales compatible with the research-based teacher quality standards.

The logic behind selecting the polytomous scale with the frequency levels as a response scale was based on the assumption that effective teachers demonstrate different or higher level of frequency of performance on the teacher quality indicators as compared to those who are less effective or ineffective teachers and who demonstrate lower level of the frequency on the same kinds of teacher quality indicators. The response scales ranged from lowest to highest level of frequency of teacher quality such as Never, Rarely, Sometimes, Often, or Always.

Open-ended Question

One open-ended question was also included in the SITE. Open-ended questions are the best source of collecting qualitative data which can provide diverse information helpful for deeper understanding of the phenomenon being studied (Geer, 1998). Open-ended questions also provide important information helpful for future research. The purpose of the single open-ended question was to take teachers' comments and suggestions, and provide them space so that they could generate new ideas more openly to provide information about their feelings about anything related to teacher evaluation or the SITE.

Selecting Predictor Variables

Two types of predictors were identified in the study: (a) teacher evaluation variables, and (b) personal characteristics variables. The rationale for their selection is provided:

a) Based on the literature review, all the six teacher evaluation variables (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development) are highly research-based and demonstrate correlation with student achievement (Gallagher, 2004; Heneman et al., 2006; Kimball, 2004; White, 2004); therefore, they may significantly predict student achievement in the Pakistani context.

b) In Pakistan, where girls and boys are provided with separate public education with the same teacher gender, it would be interesting to study the predictability of gender for student achievement. Years of teacher experience has not been found to be significantly correlated with student achievement in the United States (Gallagher, 2004; Milanowski, 2004; White, 2004). In Pakistan, however, teacher experience has provided evidence of bifurcation among Secondary School Teachers (SSTs) in regard

to their classroom performance on certain indicators (Almani, 2002; Aziz, 2010). Teacher experience, therefore, may predict student achievement in English and/or mathematics.

Expert Review and Final Refinement of Survey Instrument

After developing the initial level 44 items and response scales grouped into 6 domains, the researcher formed them into a Self-assessment Instrument for Teacher Evaluation (SITE) and included a demographics section in the beginning of the SITE. The demographics comprised several variables including teacher experience and teacher gender. After that, the researcher approached validity by including experts' and practitioners' opinions about the content of the instrument accordingly. At this stage, two panels reviewed the pilot instrument. One expert panel comprised three professors of education who had more than 20 years of teaching experience in the field of teacher education and/or testing included (1) Dr. Sally Zepeda—the major professor, (2) Dr. Allan Cohen—the methodology professor, and (3) Dr. Steve Cramer—a methodological consultant and the director of the Georgia Center for Assessment.

The second panel comprised five practitioners—Secondary School Teachers (SSTs) of mathematics or English in a public high school in Pakistan—who had varying levels of teaching experience, from 5 to 20 years. The expert panel determined if the items were clear and correctly grouped into the domains, or if the items were poorly worded or superfluous. The practitioners' panel was asked to review the items and determine if the items were clear and understandable to them, and if the items fitted with the Pakistani context. Side by side, the researcher had several 30 minutes discussion sessions with Dr. Cramer, the major professor, and the methodology professor, and a 40-50 minute web conference with the five practitioners in Pakistan.

The expert and practitioners panels reviewed the SITE and brought up some of the issues associated with the instrument. Both the panels gave comprehensive feedback and opinion on the validity of the content, relevancy of the items to the certain domains, and redundancy between the items. The dissertation committee was unanimously in agreement that the content was valid and the items were measuring what they were supposed to measure. The methodology professor suggested a couple of changes in the layout of the SITE to improve the readability and appropriateness.

In light of the critique sessions and feedback of the experts and practitioners, the researcher eliminated some of the items which were redundant or not clear to the reader, added a couple of items, moved a couple of items from one domain to another, and made minor editorial changes in the items. As a result of those modifications, the SITE was reduced to 41 items. The researcher, once again, submitted the modified version of the SITE to the dissertation committee members, as well as other experts and practitioners to get their final feedback. Finding no issues attached with the SITE, the researcher submitted the instrument to the Institutional Review Board (IRB) at the University of Georgia for approval (see Appendix B).

Pilot Study

The pilot study was conducted during August 2012. After getting the IRB approval in July 2012, the researcher sent the Self-Assessment Instrument for Teacher Evaluation (SITE) to one of his colleagues in Pakistan, who completed the Collaborative Institutional Training Initiative (CITI) training and served as a recruiter in the pilot testing as well as the data collection process of the final study. The purpose of this pilot testing was to understand whether English or mathematics teachers in public high schools in Pakistan understood the items in the SITE correctly or if they required further clarifications or modifications in the items. The

researcher emailed the SITE to the recruiter along with the guidelines comprising the complete procedure of the data collection as approved by the IRB at the University of Georgia.

For the pilot testing, the recruiter personally visited one high school and received authorization from the headmaster to administer the SITE in that school. After getting the headmaster's permission (see Appendix C), the recruiter met six teachers—3 English teachers and 3 mathematics teachers—in the school and asked them if they were interested in taking part in the study. The recruiter used the verbal script, approved by the IRB, for seeking teachers' interest in participating in the study. The verbal script in English is in Appendix D, while Urdu translation of verbal script is in Appendix E. After each teacher showed his interest in the study, the recruiter gave a consent form to each teacher to sign (see Appendix F). The recruiter also gave a copy of the consent form to each teacher for his record. After that, the recruiter distributed the SITE to each teacher. The teachers completed the SITE and handed-over the completed SITE to the recruiter who emailed the scanned copies of those questionnaires to the researcher.

Since the sample size was too small to analyze the data, the researcher interviewed those six teachers via telephone, while they were holding the SITE in their hands, and asked them if the items were clear to them, and if they were able to understand the items, and/or if they required modifications in the items. The teachers were completely satisfied with the language and the content of the SITE. The teachers agreed that the items were relevant to their context, they were able to understand the items clearly, they did not require further assistance to understand the items, and they did not suggest modifications to any items. Since the purpose of the pilot testing was met through these interviews, the researcher did not suggest further modifications in the SITE and used it for data collection.

Study Population and Sample

The population of this study involved all those teachers who taught English or mathematics to 10th graders in district Okara during 2011-2012. A majority of the public high school in Pakistan include grade 6 through 10. In rural areas where there is a shortage of primary or elementary schools, high schools take responsibility of teaching to the primary and elementary levels as well. In high schools, the Secondary School Teachers (SSTs) and Elementary School Teachers (ETSs) can also be assigned to teach high school classes. The population of this study, therefore, included SSTs as well as ESTs.

The sample of the study was selected through a convenience sampling technique. Due to the logistics, such as lack of the teachers' access to information technology and internet, and the lack of the teachers' knowledge of using the internet, it was not possible for the researcher to collect data electronically. The researcher received assistance from a recruiter in collecting data. The recruiter collected data from 155 teachers, scanned all data, and emailed them to the researcher. Fifteen teachers declined to participate in the study. The response rate was 91%.

Student Achievement Scores

Another type of instrument used for collecting student achievement data were the tests of English and mathematics given to the 10th graders by the Board of Intermediate and Secondary Education (BISE) Lahore during the annual exams conducted in March-April 2012. The BISE Lahore is the responsible body for conducting exams for secondary (grades 9-10) and higher secondary classes (grades11-12) among schools and colleges in five districts named (1) Lahore, (2) Sheikhupura, (3) Kasur, (4) Nankana Sahib, and (5) Okara.

The researcher had a detailed telephone conversation with S. A. Sajid—a District Education Officer (DEO) in Lahore, and who has been serving as a paper setter for the BISE

Lahore for many years—to understand the process of test development used by the Lahore Board. According to him, the Inter Board Committee of Chairmen (IBCC) is a governing body that coordinates activities of the various education boards related to the uniformity of academic, evaluation, and curricular standards (S. A. Sajid, personal communication, August 21, 2012). The IBCC decides about the pattern as well as the type of the questions (objective, subjective and short answer), and percent of portion of each type of question for the tests for grades 9-12 across all subjects. The recommendations of the IBCC about how the tests will be developed are sent to a Subject Selection Committee (SSC) that works within each education board.

The Subject Selection Committee in the Lahore Board meets twice in a year, selects three educationists (paper setters), initially, for each subject and sends their names to the Chairman as well as the Controller of the Examination of the Lahore Board for final approval. The initial selection of the paper setters is made on the criteria such as teacher's qualification, experience in test development, and reputation to make sure secrecy of the tests. The Chairman and the Controller Examination of the Lahore Board select, unanimously, one paper setter for each subject and invite them to develop three sets of question papers for each subject. These paper setters are provided with question papers of previous years as well as the paper patterns decided by the IBCC. The paper setters follow the IBCC recommendations as well as the table of specifications' requirements, develop three question papers across each subject, seal them, and submit to the relevant department that deals with the security and printing of the question papers. The secrecy branch publishes multiple question papers across each subject and only the secrecy branch knows which question papers are distributed among the candidates on the exam day.

Almost two weeks before the commencement of the exams, the Lahore Board's administration teams handover the sealed envelopes of question papers to the managers of the

various branches of the banks. Banks ensure confidentiality and keep the sealed question papers in lockers. On the exam day, the Superintendent of each Examination Center visits the particular bank, shows an authorization letter—issued by the Controller Examination of the BISE Lahore—to the bank manager to get the sealed envelope(s) of the question papers, brings envelope(s) to the examination hall, opens the envelope(s) in front of the invigilation staff and the candidates, and supervises the question papers distribution process among the candidates. After the students are done with the tests, the answer sheets of the candidates are collected, sealed into the bundles, and submitted to the particular bank where the Lahore Board's team collects the bundles and takes them to the BISE office, Lahore. The board officials allot fictitious roll number to each answer sheet to ensure confidentiality.

Paper marking is the second process of examination. In each district, the BISE Lahore has established cluster centers for paper marking. Qualified and expert teachers across each subject, selected through proper channels, are invited to the cluster centers where invigilation teams of the Board distribute rubrics among the teachers and give them training about how to mark the answer sheets. For each subject, the teachers are grouped into smaller teams which are headed by the subject specialists or the senior headmasters or headmistresses of the high schools, and evaluate answer sheets of the students according to the directions given in the rubrics. The heads of the groups reevaluate, randomly, some portion of the total answer sheets (probably 20% of the total answer sheets) evaluated by each teacher and guide the teachers about improving the evaluation process, if required. Based on these measures, all the answer sheets of the candidates are evaluated and result sheets are prepared and submitted to the Board's officials. The Board's administration prepares computerized results of candidates and then publishes these results in the gazette—a book—or prepare CDs of the results for public information purposes.

Data Collection

The data were collected in August and September 2012. The recruiter visited 34 public high schools in 1 district, Okara; the schools which he could conveniently visit. The recruiter personally visited each high school, met the school administrator, and received authorization from him or her to distribute the SITE among English or mathematics teachers in the school. After getting authorization from the head of the school, the recruiter met teachers, as guided by the head of the school, who taught English or mathematics to 10th graders during the academic year of 2011-2012. The recruiter, using the verbal script approved by the IRB, talked to each teacher about the project and asked about teacher's interest in the project.

After the teacher showed interest in the project, the recruiter distributed a consent letter to each teacher. The teacher read the consent form, put signature on that form, and returned it to the recruiter. The recruiter also gave a copy of that consent form to each teacher for the teacher's record. The recruiter then distributed the Self-assessment Instrument for Teacher Evaluation (SITE) to each teacher. After the teacher had completed the SITE, the recruiter collected the SITE from the teacher and placed it in the envelope.

Following the same procedure, the recruiter visited 170 teachers—101 male and 69 female—in district Okara. Out of 170 teachers, 94 males and 61 female teachers responded, while others (7 male and 8 female teachers) declined to participate in the study. The response rate was 91%. Additionally, the recruiter collected students' achievement scores in English or mathematics from each teacher (n=6570). After the data were collected, the recruiter emailed the scanned data to the researcher. The researcher carefully entered the data into SPSS, recheck to make sure the data are correctly entered, and analyzed accordingly.

Data Preparation

After the data were received, the researcher printed all the data sheets and saved them in the record. No information about the demographics was missing in the data. The researcher coded the demographics using a common coding scheme. The first step in data preparation was a necessary recoding. The recodes included only one reversed item. No missing values were found in the data. Also, no respondent provided more than one response for any item. However, a vast majority of the respondents declined to answer the open-ended question which was about their comments for anything related to the instrument or teacher evaluation. Finding the data clean, the researcher entered the data into 17.0 SPSS for data tabulation and data analysis. The researcher carefully entered each value of the data given in the questionnaires and rechecked to make sure every value was correctly entered into the SPSS. Further, the researcher created the scale scores for the six teacher evaluation predictor variables. The variables were labeled as scales.

After the teacher evaluation data were entered into SPSS, the researcher entered students' achievement scores in English or mathematics into SPSS, calculated the mean scores of each class, and used mean scores to correlate with the teacher evaluation scores on predictor variables. The descriptive results in Table 13 show that out of 155 teachers, 94 (61%) were male and 61 (39%) were female. A majority of the teachers were Secondary School Teachers (77%), particularly appointed to teach secondary classes. Years of teacher's experience ranged from 1 to 36 with the mean of 12.28. A vast majority of the teachers (87%) had a Master's Degree (academic degree) in some subject; only 13% had a Bachelor's Degree.

All of the sampled teachers held at least one professional teaching degree; more than half of them (58%) had a Master of Education Degree (M.Ed.) as the highest professional teaching degree, as compared to the other teachers (38%) who held Bachelor of Education (B.Ed.)

Degrees. Class size varied from as minimum as 5 students to a maximum of 77 students, with the mean number of students more than 40 students ($\bar{x}$ =42.34). After the data were entered into SPSS and the frequencies were run to find the missing values, the data were ready to be analyzed. The researcher assigned the scales for the six teacher evaluation predictors.

Table 13

Demographic and Raw Response Rate Description of the Respondents (N=155)

Variable			Value
Gender			
Male		n=94	61%
Female		n=61	39%
Subject Taught			
English		n=81	52%
Male		n=48	59%
Female		n=33	41%
Mathematics		n=74	48%
Male		n=46	62%
Female		n=28	38%
Job Title			
Elementary School Teacher(EST/SV/etc)		n=33	22%
Secondary School Teacher (SST)		n=122	78%
Highest Academic Degree			
Bachelor's		n=21	13%
Master's		n=134	87%
Highest Professional Degree			
C.T. (Certificate of Teaching)		n=5	4%
B.Ed.		n=60	38%
M.Ed.		n=90	58%
Years of Experience	Min=1	Max=36	Mean=12.28
Class Size	Min=5	Max=77	Mean=42.28

At this stage, Cronbach Alpha was calculated to measure the reliability of the teacher evaluation scale. The results are given in Table 14. Since the purpose was to develop an overall reliable teacher evaluation instrument, the researcher calculated the internal reliability with all 41 items of the instrument. The SITE demonstrated a relatively high level of overall Cronbach alpha reliability (α =.86).

Table 14

Distributions and Reliability of Predictor Scale Variables

Scales	Number of Items	M	SD	Mean Item Means	Cronbach's Alpha
Subject Matter Knowledge	6	25.71	2.203	4.29	.62
Instructional Planning and Strategies	7	29.70	2.561	4.24	.60
Assessment	6	24.44	2.735	4.07	.61
Learning environment	7	30.15	2.655	4.30	.71
Effective communication	7	28.74	2.708	4.10	.61
Continuous Professional Development	7	30.11	2.900	4.30	.66

At the next stage, the researcher calculated the reliability for each of the six teacher evaluation subscales separately. In examining the reliability of the subscales, an item (item 18) contributing to the variable Assessment appeared to be a problematic. A closer examination of the item "my high achieving students evaluate class tests of their class-fellows" showed that the respondents might have become confused in the intent of the item. They might have thought that

effective teachers should or should not take help of high achieving students to assess their class-fellows' tests; so they might have become biased over this item.

The subscale, Assessment, showed an internal reliability value of .51 when item 18 was included in the reliability analysis; excluding this item, the reliability reached to .61 which was similar to the reliability level of other subscales. Since this item did not contribute to the total analysis of the subscale, and the study employed the additive indices in the methodology where the reliability is not an issue (Ranne, 2011), the researcher eliminated this item and revised the scale with the remaining six items to use for further analysis. Table 14 shows that the reliability of the subscales ranged from .60 to .71 which is reasonably high for a self-assessment instrument where the issue of conflict of interest for teachers may arise when they assume that their evaluation scores on teacher effectiveness instrument would be correlated with the achievement of their students, especially on high-stake tests.

In the following, Figures 3 to 8 show the distribution of each scale.

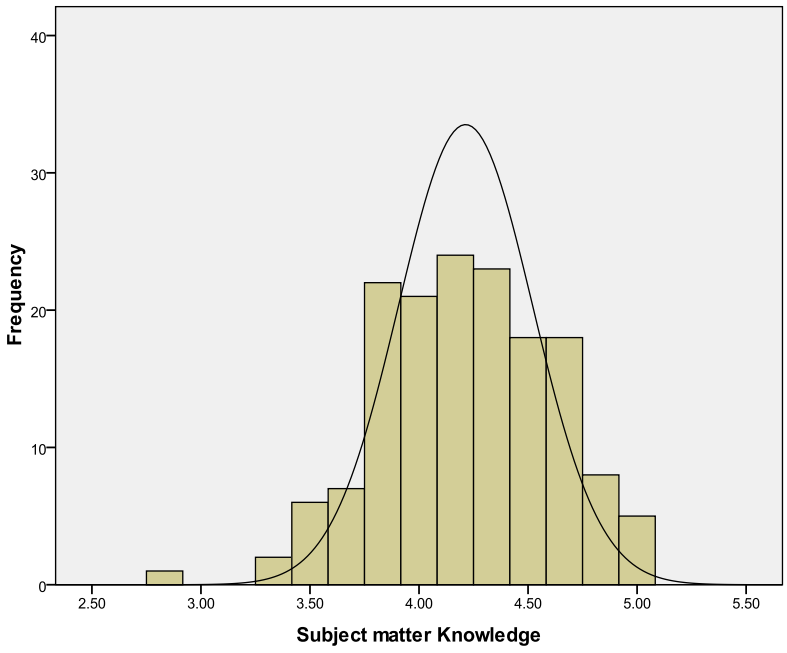

Figure 3. Distribution of Subject Matter Knowledge Scale

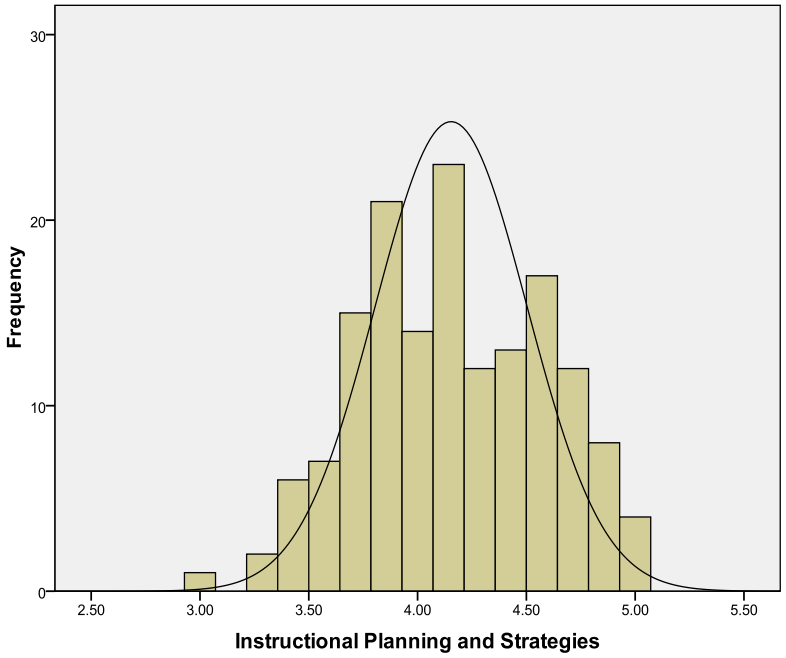

Figure 4. *Distribution of Instructional Planning and Strategies Scale*

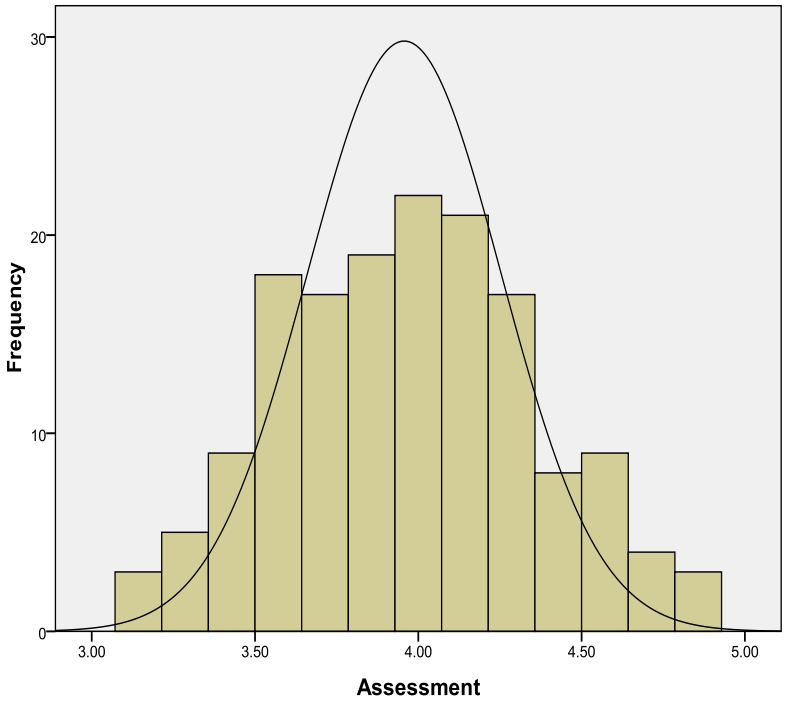

Figure 5. Distribution of Assessment Scale

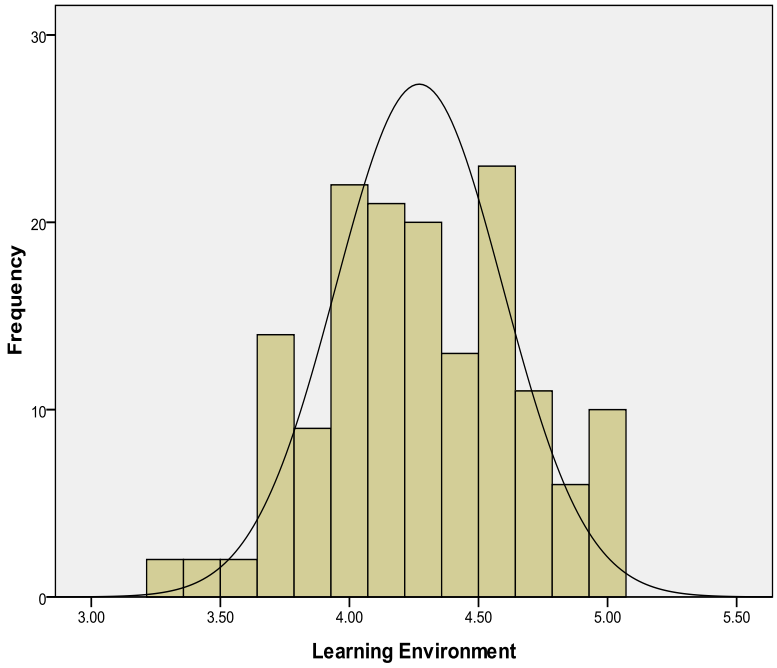

Figure 6. Distribution of Learning Environment Scale

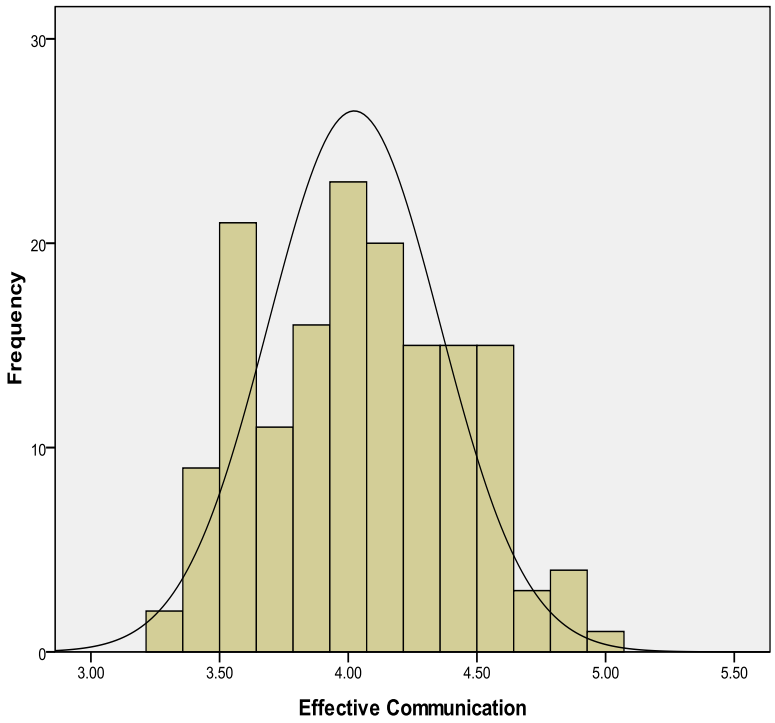

Figure 7. Distribution of Effective Communication Scale

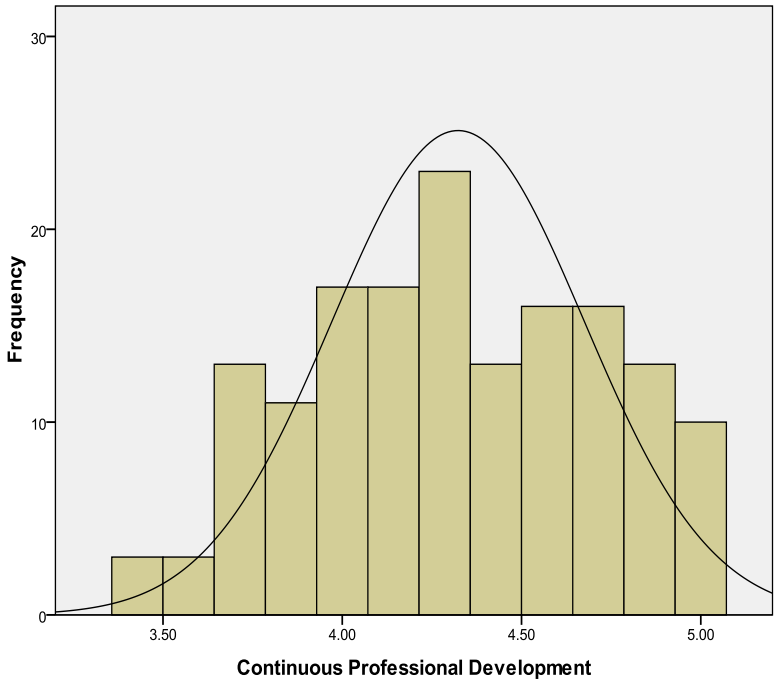

Figure 8. Distribution of Continuous Professional Development Scale

After the reliabilities were calculated, the researcher calculated the relationships between the six variables of the teacher evaluation instrument. The purpose of measuring the relationships between these variables was to ascertain that the six constructs were discriminant from each other, and they avoided substantive redundancy (Messick, 1989). The results showed that all the components of teacher evaluation had lower to moderate significant relationship with each other. A summary of the relationships is shown in Table 15.

The highest significant relationship was found between Continuous Professional Development and Effective Communication, r=.58, p <.01. The lowest significant relationship was found between Professional Development and Learning Environment, r=.19, p <.05. The significant intercorrelations between teacher evaluation components provided direction for further analyses related to the three research questions.

Table 15

Intercorrelations of Component Measures of Teacher Evaluation (N=155)

Components	1	2	3	4	5	6
Subject Matter Knowledge	1.00					
Instructional Planning and Strategies	48**	1.00				
Assessment	.27**	.26**	1.00			
Learning Environment	.26**	.43**	.22**	1.00		
Effective Communication	.40**	.38**	.38**	.39**	1.00	
Continuous Professional Development	.42**	.42**	.42**	.19*	.58**	1.00

**$p<0.05$ (2-tailed), *$p<0.01$ (2-tailed)

Data Analysis

The data analysis was conducted using SPSS 17.0. The quantitative analyses comprised correlations, simple and multiple regressions, and t-tests for independent samples. A data analysis method for each research question is described in the following discussion.

For research question 1: To what extent To what extent do six performance evaluation scales (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development) measured through a self-assessment instrument separately predict student performance in English or mathematics in Pakistan was analyzed to determine what predicts the observed variance in the students' achievement in English or mathematics for each component. A series of bivariate analyses were conducted to determine the relationships. Simple regressions were performed on

six teacher evaluation predictors for English as well as Mathematics separately and Correlation Coefficient of Determination (r^2) was calculated.

For research question 2: To what extent do the six scales measured through a self-assessment instrument combine to predict student performance in English or mathematics in Pakistan, a series of bivariate analyses were conducted to determine the correlations. Multiple regression analysis was run with the six teacher evaluation predictors taking English or mathematics as a dependent variable. Since the predictor variables were continuous, Pearson Correlation Coefficients and Coefficient of Determination (r^2) were calculated to investigate the relationships between the predictor variables and student achievement in English or mathematics.

For research question 3: Does the addition of teacher gender and teaching experience to the multiple regression model significantly increase the value of prediction in English or mathematics in Pakistan, multiple regression analysis along with stepwise, forward, backward, or enter method was conducted combining the teacher gender and experience with the six teacher evaluation predictors. A t-test for independent samples was also conducted to confirm the significant relationship of gender with other significant predictors of teacher evaluation and student achievement in English.

Limitations

This study involves some of the limitations. First, the study did not involve a random sampling technique which is the preferable way to obtain a representative sample and is important for inferential statistics (Fraenkel & Wallen, 2009). In non-random sampling (i.e., convenience sampling), each individual has not equal and independent chance of being selected for the sample. Also, the number of male (n=91) and female (n=64) teachers was not equal. Similarly, the number of teachers selected from the urban schools (n=117) was much higher than

the teachers selected from the rural schools (*n*=38). Due to the non-representative sampling technique, the results could be interpreted as biased.

The second limitation of the study was associated with a relatively a small sample size (*n*=155). Due to limited resources, the recruiter was not able to collect data from more teachers. Especially, in the Pakistani context there is not coeducation grouping in public schools, and culturally it was complicated to get more data from female teachers.

Thirdly, the Self-assessment Instrument for Teacher Evaluation (SITE) was distributed in the English language and not translated into Urdu language which is the national language of Pakistan. Though Pakistan had been a British Colony where English had been used as a medium of instruction for decades, there may be some teachers who understood meanings of some of the items differently despite their high qualifications with the English language. Due to these limitations, any generalizations should be made with caution.

CHAPTER 4

RESEARCH FINDINGS

The purpose of this study was to measure the relationship between teacher evaluation scores and 10th graders' achievement in English or mathematics in Pakistan. This chapter reports the findings in regard to the following overall research questions:

1. To what extent do six performance evaluation scales (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development) measured through a self-assessment instrument separately predict student performance in English or mathematics in Pakistan?
2. To what extent do the six scales measured through a self-assessment instrument combine to predict student performance in English or mathematics in Pakistan?
3. Does the addition of teacher gender and teaching experience to the multiple regression model significantly increase the value of prediction in English or mathematics in Pakistan?

Findings in Regard to Research Question # 1

The first research question asked "To what extent do six performance evaluation scales (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development) measured through a self-assessment instrument separately predict student performance in English or mathematics in Pakistan?" To answer this research question, the researcher ran the scatter plots

to investigate the outliers or influential points that could affect the results of the study. The scatter plots identified a couple of cases as outliers for English as well as mathematics. Since these cases were not influential points, the researcher did not remove them and ran the correlation analyses including those cases. Table 16 gives the summary results of the correlations between teacher evaluation scales and student achievement.

Table 16

Relationship Between Teacher Evaluation and Student Achievement in English and Mathematics

Teacher Evaluation Variables	English	Mathematics
Subject Matter Knowledge	.43*	.13
Instructional Planning and Strategies	.45*	-.15
Assessment	.24*	-.07
Learning Environment	.21	.16
Effective Communication	.33*	.11
Continuous Professional Development	.41*	.00

*$p < .05$

The first research question was analyzed in two parts. In first part, the relationship between teacher evaluation scores and student achievement in English was investigated. The results in Table 16 show that 5 of 6 scales of teacher evaluation were significantly correlated with student achievement in English. Instructional Planning and Strategies scale showed the highest correlation with student achievement in English (.45), followed by Subject Matter Knowledge (.43), and Continuous Professional Development (.41). Learning Environment, however, did not show significant relationship with student achievement in English (.21). The

Instructional Planning and Strategies scale independently explained 20% of the variance in student achievement in English, followed by the Subject Matter Knowledge which explained almost 18% variance in student achievement in English. Learning Environment explained only 5% of the total variance in student achievement in English.

The second part of the first research question involved predicting student achievement in mathematics with six teacher evaluation predictors. All the correlations were nonsignificant, demonstrating no significant relationship with student achievement in mathematics.

Findings in Regard to Research Question # 2

The second research question asked "To what extent do the six scales measured through a self-assessment instrument combine to predict student performance in English or mathematics in Pakistan?" This research question was analyzed at two stages. Initially, a multiple regression was run to determine which of the teacher evaluation variables, when combined, explained the variance in student achievement scores in English. To determine this outcome, all six teacher evaluation variables were entered together into the multiple regression dialog box and English was entered into the dependent variable window. Selecting the stepwise regression method from the Method drop-down box, the researcher ran the multiple regression analysis. See Table 17.

Table 17

Collective Teacher Evaluation Model for Student Achievement in English

Parameter	Parameter Estimate (b)	Standardized Estimate (Beta)	t	p
Subject Matter Knowledge	7.331	.292	2.751	<.05
Instructional Planning and Strategies	9.215	.335	3.069	<.05

When all six independent variables were employed together, only two of them, Subject Matter Knowledge and Instructional Planning and Strategies, significantly predicted student achievement in English, $R^2 = .27$, F (2, 78) = 14.345, $p < .001$. The r^2 value in the best collective teacher evaluation model showed that 27% of the observed variance in student achievement in English could be explained through Subject Matter Knowledge and Instructional Planning and Strategies. The results indicated that the students of teachers with higher teacher evaluation scores on Subject Matter Knowledge and Instructional Planning and Strategies were expected to have higher achievement in English. The rest of the predictors did not significantly contribute to the student achievement in English.

At the second stage, a similar procedure was adopted for measuring the prediction value of the six teacher evaluation variables for student achievement in mathematics. As shown previously in Table 16, all six teacher evaluation variables were individually nonsignificant with student achievement in mathematics; however, a multiple regression was run to determine which of the teacher evaluation variables combined would explain the variance in student achievement in mathematics. For this purpose, the researcher entered all six teacher evaluation variables into multiple regression dialog box, whereas student achievement in mathematics was entered into the dependent variable window. Selecting the stepwise regression method from the Method drop-down box, the researcher ran the multiple regression analysis. The SPSS output showed a statement indicating "no variables were entered into the equation." It was, perhaps, due to the non-significant relationships of teacher scores with student achievement in mathematics, as revealed in Table 16. The researcher, then, used the backward method and reran the regression analysis.

The results showed that Subject Matter Knowledge, Instructional Planning and Strategies, and Learning Environment predicted 13% of the observed variance in student achievement in mathematics; however, Learning Environment showed non-significant relationship with student achievement. Lastly, the researcher used the Enter method and ran the regression analyses with all possible combinations of the predictors. Table 18 shows that only two predictors—Subject Matter Knowledge and Instructional Planning and Strategies—significantly predicted 9% of observed variance in student achievement scores in mathematics, $F(2, 71) = 3.586$, $p = .033$.

Table 18

Collective Teacher Evaluation Model for Student Achievement in Mathematics

Parameter	Parameter Estimate (b)	Standardized Estimate (Beta)	t	p
Subject Matter Knowledge	9.254	.326	2.349	<.05
Instructional Planning and Strategies	-8.349	-.334	-2.409	<.05

The Subject Matter Knowledge had significant positive regression weights, indicating teachers with high scores on this scale were expected to have higher student achievement in mathematics. However, the Instructional Planning and Strategies variable showed negative regression weights, indicating that after accounting for Subject Matter Knowledge scores, those teachers with higher scores on Instructional Planning and Strategies were expected to demonstrate lower student achievement in mathematics. In such a case, the variable Instructional Planning and Strategies played the role of suppressor or mediator variable, indicating that the variable Instructional Planning and Strategies is uncorrelated or relatively unrelated with the criterion (student achievement) but is related to other predictors such as Subject Matter

Knowledge and significantly predicts student achievement indirectly through Subject Matter Knowledge. The rest of the four independent variables did not contribute to the student achievement in mathematics.

Findings in Regard to Research Question # 3

Third research question asked "Does the addition of teacher gender and teaching experience to the multiple regression model significantly increase the value of prediction in English or mathematics in Pakistan?" Gender and teacher experience were demographic independent variables and named as variables of personal characteristics. This research question was also analyzed in two parts. Initially, the researcher entered the best fitted model that included two significant teacher evaluation variables (i.e., Subject Matter Knowledge and Instructional Planning and Strategies) into the multiple regression dialog box and added teacher gender and teacher experience with that model, whereas student achievement in English was entered into the dependent variable window. Stepwise method was used to run multiple regression analysis. The results of the analysis are shown in Table 19.

Table 19

Final Teacher Evaluation Model for Student Achievement in English

Parameter	Parameter Estimate (b)	Standardized Estimate (Beta)	t	p
Subject Matter Knowledge	8.056	.321	3.090	<.05
Instructional Planning and Strategies	6.866	.242	2.219	<.05
Gender	4.634	.231	2.317	<.05

The results showed that gender contributed 5% of observed variance in student achievement scores in English by increasing the teacher evaluation prediction from 27% to 32%. Based on this final model, 32% of the total variance in student achievement in English can be explained by teachers' scores on Subject Matter Knowledge, Instructional Planning and Strategies, and teacher gender, $R^2 = .32$, $F(3, 77) = 11.888$, $p < .001$. A following test—t-test for independent samples—was also run on gender to measure its significance with student achievement in English. The results showed that female teachers demonstrated significantly higher mean scores (M= 40.20, SD = 10.55) than male teachers (M= 33.93, SD= 8.67) in English, $t(79) = -2.925$, $p=.004$. Teacher experience, on the other hand, did not contribute to student achievement in English.

In the second part, the researcher followed the same procedure for mathematics, entered Subject Matter Knowledge and Instructional Planning and Strategies variables into the multiple regression model and added teacher gender and teaching experiences into the model, entered student achievement in mathematics into the dependent variable window, and ran the multiple regression analysis using the stepwise, backward, and Enter regression methods. Stepwise method showed the same message (i.e., no variables were entered into the equation). The backward and Enter methods showed similar findings, indicating that neither the gender nor the teacher experience added variance in student achievement scores in mathematics. So the final prediction model for student achievement in mathematics remained unchanged. Summarizing the results, the study found positive relationships between teacher evaluation scores and student achievement in English, and mixed—positive as well as negative—results with student achievement in mathematics. Teacher gender also significantly contributed to the student achievement in English.

One open-ended question was also asked in the SITE to gather comments and insights on anything related to teacher evaluation. However, a vast majority of the teachers did not respond to the open-ended question. The lack of response was, probably related to language proficiency because SITE was in English and the teachers more than likely did not feel comfortable with writing in English. There was one stark pattern in who responded to the open-ended question in that only male teachers of mathematics responded; no female teachers responded. A couple of male teachers indicated that the SITE "questionnaire is organized" or "the questionnaire is good for improving learning environment" or "the questionnaire increased my knowledge." One teacher commented that "good relations with teachers, head teachers, and parents are needed." A couple of male teachers, however, complained about the lengthy course of studies needed to teach mathematics. One teacher commented:

> I have studies all the categories (items); they are suitable for improving the educational system. The new curriculum of mathematics is so much lengthy. Finally, I want to say that the schools should be given equal resources, and (then) the results would be good.

Another mathematics teacher commented, "It is difficult to teach math…usually cramming method is used. I try my best to teach math in logical ways. I want students use mathematical knowledge in daily life."

Based on these comments, it can be assumed that these teachers liked the SITE and they found it helpful for the improvement of the learning environment. Further qualitative studies about teachers' perspectives, especially mathematics teachers' perspectives, of the SITE, teacher evaluation, or mathematics curriculum would be highly beneficial.

It is important to understand the detailed discussion of the findings along with the implications for policy, practice, and research. The discussion provides the policymakers an understanding of the effectiveness of the National Professional Standards; the district education

authorities and school administrators would also be able to identify effective teachers and place them accordingly. Chapter 5 provides a detailed discussion of these findings and implications. Recommendations for future research are also given in Chapter 5.

CHAPTER 5

DISCUSSION OF FINDINGS

The purpose of this study was to measure the relationship between teacher evaluation scores and 10th graders achievement in English or mathematics in the 2012 annual examination conducted by the Board of Intermediate and Secondary Education (BISE) Lahore, Pakistan. A quantitative approach was used to measure the relationship between teacher evaluation scores and student achievement. Chapter 5 presents an overview of the study, a summary of the findings, principle findings, discussion of the findings, and the implications for policy, practices, and future research.

Overview of the Study

The purpose of this study was to measure the relationship between teacher evaluation scores and 10th graders' achievement in English or mathematics in the 2012 annual exam conducted by the Board of Intermediate and Secondary Education (BISE) Lahore, Pakistan. A second purpose of this study was to develop a valid and a reliable self-assessment instrument for teachers in the US, providing them an alternative of evaluators' ratings which have traditionally been proven biased and flawed (Heneman et al., 2006; Kauchak et al., 1985; Milanowski, 2004; Peterson, 2000). The population of the study involved public high school teachers in Pakistan who taught English or mathematics to 10th graders in the academic year 2011-2012.

A Self-assessment Instrument for Teacher Evaluation (SITE) was developed to measure teachers' evaluation scores on six teacher evaluation predictors namely: Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective

Communication, and Continuous Professional Development. Teacher gender and teacher experience were selected as teacher characteristic variables and used to investigate whether they contributed to the variance in student achievement in English or mathematics. The six teacher evaluation variables were research-based teacher quality indicators as described by Danielson (1996) and Marzano (2010), and summarized by Stronge (2010). These teacher evaluation variables were also part of the National Professional Standards for Pakistani public school teachers designed by the Ministry of Education (2009), Pakistan.

The Self-assessment Instrument for Teacher Evaluation (SITE) was developed based on the teacher quality indicators as sampled by Stronge (2010). The process of the SITE development is explained in Chapter 3. The SITE is comprised of 41 items and each item is comprised of five response levels ranging from "Never" to "Always". The SITE was developed and content validated with the help of an expert panel as well as a panel of practitioners. The expert panel included senior professors of Educational Administration and Policy and Educational Testing at the University of Georgia; the practitioners were public school teachers in Pakistan. The SITE was developed, modified in the light of the feedback of the panels, pilot tested, and used for data collection.

The study was conducted in 34 public high schools in district Okara, province Punjab, Pakistan. The data were collected from 155 teachers; 91 of them were male and 61 were female. The achievement scores of those students who were taught English or mathematics by these teachers were also collected from the teachers. Based on these scores, the results were analyzed according to the methodology described with detail in Chapter 3. A summary of the finding is presented here.

Summary of the Findings

The following research questions guided the study:

1. To what extent do six performance evaluation scales (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development) measured through a self-assessment instrument separately predict student performance in English or mathematics in Pakistan?
2. To what extent do the six scales measured through a self-assessment instrument combine to predict student performance in English or mathematics in Pakistan?
3. Does the addition of teacher gender and teaching experience to the multiple regression model significantly increase the value of prediction in English or mathematics in Pakistan?

The first research question asked, "To what extent do six performance evaluation scales (Subject Matter Knowledge, Instructional Planning and Strategies, Assessment, Learning Environment, Effective Communication, and Continuous Professional Development) measured through a self-assessment instrument separately predict student performance in English or mathematics in Pakistan?" Six teacher evaluation scales identified in the literature as additive components of the teacher effectiveness construct were analyzed. The first scale (Subject Matter Knowledge) and the third scale (Assessment) included six items each; the rest of the four scales included seven items each. The mean score of the 6 scales ranged from as low as 24.44 (Assessment) to as high as 30.11 (Continuous Professional Development).

The first research question was analyzed in two parts: (1) prediction of English achievement, and (2) prediction of mathematics achievement. In part one, simple correlation

analyses were performed separately on each of the six independent variables, taking the student achievement in English as dependent variable separately. The results revealed five out of six teacher evaluation variables were significantly correlated with student achievement in English. The correlation coefficients ranged from as low as .21 to as high as .45. The Instructional Planning and Strategies was the strongest explanatory variable which explained almost 20% of the variance in student achievement in English, followed by Subject Matter Knowledge which explained almost 18% of the variance, and Continuous Professional Development with 17% of the variance in student achievement in English.

In the second part, simple linear regression analyses were performed separately for the six teacher evaluation variables, with student achievement in mathematics as a dependent variable. The correlation coefficients ranged from .16 (Learning Environment) to -.15 (Instructional Planning and Strategies). The results revealed that none of the teacher evaluation variables was significantly correlated with student achievement in mathematics. The results also revealed that Instructional Planning and Strategies was negatively correlated with student achievement in mathematics, followed by the Assessment with negative (-.07) correlation coefficient.

The second research question asked, "To what extent do the six scales measured through a self-assessment instrument combine to predict student performance in English or mathematics in Pakistan? The second research question was analyzed in two parts. Initially, a multiple regression analysis was performed employing all six teacher evaluation scales as predictors, with student achievement in English as outcome variable. Coefficients of correlation and coefficient of determination (r^2) were also calculated. The model fit equation was also calculated. The

results revealed that two predictors, Instructional Planning and Strategies and Subject Matter Knowledge, explained almost 27% of variance in student achievement in English.

In the second part, multiple regression analysis was conducted on the same six teacher evaluation predictors with student achievement in mathematics as an outcome variable. The coefficients of correlation and coefficient of determination (r^2) were calculated for the multiple regression analysis. The model fit equation was also calculated. The results revealed that Subject Matter Knowledge and Instructional Planning Strategies significantly predicted 9% of variance in student achievement in mathematics. The results showed that Instructional Planning and Strategies played the role of a mediator variable, indicating that the variable Instructional Planning and Strategies was uncorrelated or relatively unrelated with student achievement in mathematics, but was related to other predictor (Subject Matter Knowledge) and significantly predicted student achievement in mathematics indirectly through Subject Matter Knowledge.

The third research question asked, "Does the addition of teacher gender and teaching experience to the multiple regression model significantly increase the value of prediction in English or mathematics in Pakistan? This research question was also analyzed in two stages. At the first stage, multiple regression analysis was performed on all six teacher evaluation predictors adding teacher gender and teacher experience as predictors, with teacher student achievement in English as an outcome variable. Coefficients of correlation and coefficient of determination (r^2) were calculated. The model fit equation was also calculated. The results revealed that gender accounted for 5% of variance in student achievement in English, increasing the explained variance from 27% to 32%. Teacher experience did not contribute to student achievement in English. At the second stage, a multiple regression analysis was performed with all six predictor variables, with student achievement in mathematics as an outcome variable. The results revealed

that neither teacher gender nor teacher experience contributed to observed variance in student achievement in mathematics.

Principle Findings

There were three principle findings of this study:

1. Five of the six teacher performance evaluation predictors measured through the Self-assessment Instrument for Teacher Evaluation (SITE) significantly predicted student achievement in English but not in mathematics. These findings were consistent with the previous research (Gallagher, 2004; Heneman et al., 2006; Kimball et al., 2004; Milanowski, 2004; Odden, 2004). SITE provided valid and reliable measure of teacher evaluation. The six domains showed low to moderate intercorrelations, indicating they were not measuring the same thing, and thus avoided substantive redundancy (Messick, 1989). The overall reliability of .86 for the SITE was also relatively high.
2. Two of the six teacher evaluation variables significantly predicted student achievement in English as well as in mathematics, taking Instructional Planning and Strategies and as a mediating variable for student achievement in mathematics.
3. Teacher gender significantly increased the value of prediction in student achievement in English only. Teacher experience did not contribute to the prediction models for English and mathematics. The results of this study are partly consistent with previous research (Gallagher, 2004; White, 2004).

Discussion of the Findings

Finding 1: Five of the six teacher performance evaluation predictors measured through the Self-Assessment Instrument for Teacher Evaluation (SITE) significantly predicted student

achievement in English, independently, but not in mathematics. These findings confirmed the alignment with the findings with previous research based on multiple data sources including evaluators' ratings.

As examined in Chapter 2, the literature on the relationship between teacher evaluation and student achievement in English and mathematics can be categorized into two general categories: (a) the studies based on Marzano's Causal Teacher Evaluation model (2010), and the studies based on Danielson's Framework for Teaching (1996). Combining both models, the researcher was able to select six teacher performance variables and develop a Self-assessment Instrument for Teacher Evaluation (SITE). The SITE-based results of this study confirmed and extended the findings of those studies which employed Marzano's (2010) and Danielson's (1996) models.

Marzano (2010) found from a study conducted in Oklahoma (Phase 1) that 5 of 9 essential indicators of teacher quality significantly showed small or moderate relationship with student achievement in reading (r =.33 to .53) and mathematics (r =.31 to .39); at Phase II, however, 6 of 9 correlations were significant for reading (r =.11 to .40), and only 1 for mathematics (r =.04 to .40). The current study confirmed these previous findings as the SITE based domain-wise correlation between teacher evaluation scores showed smaller or moderate positive relationships with student achievement in English (r=.21 to .45), but no significant correlation with student achievement in mathematics (r= -.15 to .16).

Another important comparison of this study can be made with the studies based on Danielson's Framework for Teaching (1996). Studies conducted at Cincinnati, Washoe, Vaughn, and Coventry presented average correlations between teachers' overall evaluation scores and student achievement in reading and mathematics. A comparative analysis of those correlations is

presented along with the findings of this study in Table 20. The Table shows that the study of Gallagher (2004) conducted at Vaughn found the highest average correlation between teacher evaluation composite score, based on multiple data sources, and student achievement in reading (r =.37), followed by the study of Milanowski (2004) at Cincinnati (r =35).

Table 20

Comparative Analysis of the Average Correlations Between Teacher Evaluation Scores and Student Achievement in the United States and Pakistan

Sites	Data Sources	Grades	Reading/English	Mathematics
Cincinnati 3 year average:	Classroom observation, portfolio	3-8	.35	.33
Coventry 3 year average:	Observation, dialogue with teacher, portfolio	2-6	.24	-.03
Washoe 3 year average:	Self-assessment, observation, artifacts etc.	3-6	.22	.22
Vaughn 3 year average:	observation, lesson plans, student work, and others	2-5	.37	.26
Pakistan	Self-assessment (SITE)	10	.35	.03

The average correlation of the six SITE-based teacher evaluation predictors with student achievement in English was found to be positive (.35) and similar to the results found at Vaughn and Cincinnati, and higher than the results found at Coventry and Washoe. However, the average correlations between teacher evaluation scores and student achievement in mathematics at Cincinnati, Vaughn, and Washoe were found higher than the results found in Pakistan.

The study conducted by White (2004) at Coventry, Rhode Island, found essentially no correlation between teacher's overall evaluation scores and student achievement in mathematics

based on 3 year weighted average (.03), anticipating small sample size (*n*=78) as a main reason of such results. These findings are also similar to the findings of this study where the correlation between the teachers' average evaluation scores and student achievement in mathematics was found to be very small (.03), even with a smaller sample size (*n*=74).

The small or moderate positive relationships between teacher evaluation and student achievement have been found in previous studies, so they are not unexpected. For example, Heneman et al. (2006), describing smaller correlations (.11 to .22) stated:

> Note that one would not expect to find a perfect or even near-perfect correlation between evaluation scores and student achievement, given the various other factors that influence both. On the student achievement side, tests are not perfect measures of student learning, nor is teacher behavior its only cause. Teacher evaluation scores are also not perfect representations of teachers' actual classroom behavior. (p. 5)

Milanowski (2004) also gave similar arguments on relatively small correlations (.3 to .4) between teacher evaluation and student achievement by stating:

> It is important to recognize that high correlations between teacher evaluation scores and student achievement scores are unlikely to be found for reasons including error in measuring teacher performance, error in measuring student performance, lack of alignment between the curriculum taught by teachers and the student tests, and the role of student motivation and related characteristics in producing student learning. (p. 50)

Therefore, comparing the results of the current study with previous research, it can be concluded that the findings of this study confirmed that the SITE provided a consistent valid and reliable evidence of teacher quality measures.

The SITE measured what it was supposed to measure; the six domains showed less or moderate intercorrelations, ranging from .19 to .58. These intercorrelations between domains were smaller than the correlations found at Cincinnati (Milanowski, 2004) and Vaughn (Gallagher, 2004), indicating the domains or scales were not measuring the same thing, and they avoided substantive redundancy (Messick, 1989). The overall reliability of the SITE was also

relatively high (α =.86). Additionally, the teacher evaluation scores based on SITE significantly predicted student achievement in English, indicating that the SITE has *some* predictive validity of student achievement in English.

Instructional Planning and Strategies, which explained almost 20% of the variance, was the strongest explanatory variable, among other variables, of student achievement in English. The finding was not surprising because Instructional Planning and Strategies are strong evidence of teacher quality. Marzano (2010) found encouraging results based on teaching strategies conducted at Oklahoma. Almani (2002) also found similar results in Pakistan that teachers, especially female teachers, rated themselves higher on instructional planning after they received in-service training. Effective teachers are expected to use multiple instructional strategies, materials, and assessment techniques to meet students' needs and to maximize their learning (Tomlinson, 1999). Effective teachers engage, motivate, and maintain students' attention to the lesson, use the maximum time teaching, and use knowledge of available resources to determine what resources they need to support learning (Buttram, & Waters, 1997; Stronge, 2010).

Subject Matter Knowledge explained almost 18% of the observed variance in student achievement in English. Subject Matter Knowledge is a good indicator of teacher quality. This is what a teacher brings to his or her class before he or she uses any teaching strategy. If a teacher has strong knowledge of a subject matter, he/she is expected to transfer the high quality knowledge to the students (Stronge, 2002, 2007). Effective teachers demonstrate the ability to link the present with future learning experiences, use school and community resources, and to teach according to the students' intellectual, emotional, and physical development needs (Stronge 2010). Subject Matter Knowledge is equipped with the instruction provided to the pre or in-service teachers in teacher training (Baumert et al., 2010).

Continuous Professional Development, which independently explained 17% of the variance in student achievement in English, was third in importance related to the teacher effectiveness indicator. Professional development is a continuous process of maintaining a professional demeanor, participating in professional growth opportunities, and contributing to the profession (Stronge, 2010). Since it is assumed that more effective teachers will demonstrate higher level of professionalism, less effective teachers also pretend to be highly professional and can rate themselves higher on teacher evaluation indicators (Peterson, 2000). The results of this study, however, demonstrated that effective teachers produced higher student achievement, at least, in English. This finding increases the validity of the SITE as well as the fairness of teachers in evaluating their performance.

Effective Communication explained 11% of the variance in student achievement in English, followed by the *Learning Environment* which explained almost 5% of the observed variance in student achievement in English. *Effective Communication* is a process of communicating with students during instruction process in which effective teachers understand the flow of the classroom, use instructional objective plan effectively, and use effective communication to orient students and to help them integrate new information with previously learned information (Worley, Tistworth, Worley, & Cornett-DeVito, 2007). A positive classroom *Learning Environment* minimizes classroom disruption, and helps teachers to perform better. Effective teachers provide students opportunities to ask questions, and increase student interactions with their classmates as well as their teachers (Stronge, 2010).

All these teacher evaluation variables, however, did not help explain variance in student achievement in mathematics. There may be various reasons of such results. First, a majority of the teachers had at least B.Sc. Degree, indicating an equal level of subject matter knowledge.

Only 2% of the mathematics teachers had Master Degrees in mathematics. Since teachers did not differ significantly in their qualification and subject matter knowledge, the non-significant relationships between teacher evaluation scores and student achievement in mathematics were not unexpected.

Second, depending on the availability of limited resources for mathematics teachers in public schools, the teachers might not have used various teaching strategies in the classroom. Third, responding to the open-ended question given in the SITE, a couple of teachers complained about lengthy coursework in advanced studies in mathematics. It can be inferred from their responses that teachers might have not found enough time to assess students' works and tests frequently. Fourth, a majority of the schools, especially the urban schools, are overcrowded in Pakistan. It can be anticipated that mathematics teachers could not concentrate on developing interaction with students and creating an environment conducive to learning; rather, they might focused on completing the lengthy coursework well in time. Based on these reasons, the correlation between teacher evaluation scores and student achievement in mathematics were non-significant.

Finding 2: When combined, two of the six teacher evaluation variables significantly predicted student achievement in English as well as in mathematics, taking a predictor as mediating variable for student achievement in mathematics.

According to the findings, the Instructional Planning and Strategies, and Subject Matter Knowledge showed highest levels of correlations (.45 and .43) and coefficient of determination (r^2) values (20%, and 18%), respectively with student achievement in English. Other teacher evaluation variables showed, however, less relationship with student achievement in English. Collectively, only the Instructional Planning and Strategies, and Subject Matter Knowledge

combined to predict student achievement in English and explained 27% of the observed variance in student achievement in English. In the presence of various other factors such as student motivation, parents' income and education, and school resources which play a pivotal role in contributing to student achievement, the results—that 27% of the observed variance in student achievement can be explained by the Instructional Planning and Strategies and Subject Matter Knowledge—are highly encouraging.

The SITE was used for the first time in Pakistani public high schools where teachers do not have any tradition or experience in self-evaluating their performance. Teachers are evaluated on entirely different indicators based on personality characteristics given in the Performance Evaluation Report (PER) such as a teacher's intelligence, knowledge of religion, emotional stability, appearance and so on. Therefore, it is possible that the teachers might not have a deeper understanding of all the teacher evaluation variables used in the SITE.

Another important possibility of the non-significant predictors of student achievement might be teachers' understanding of how the criterion—the student achievement—will be used for correlation. It is a common practice in Pakistan that a teacher's effectiveness is judged by his or her students' pass percentage, and not their actual achievement. Teachers might have evaluated themselves on the SITE keeping in mind that their students' pass percentage rate would be correlated with teachers' scores. It is possible, therefore, that teachers showed higher pass percentage but lower student achievement and mean scores in English or mathematics.

Finding 3: Teacher gender significantly increased the value of prediction in student achievement in English only. Teacher experience did not contribute to the final prediction model for English as well as mathematics. The results of this study are partly consistent with previous research (Gallagher, 2004; White, 2004).

Teacher gender was a significant predictor of student achievement in English. Gender added 5% of the observed variance in student achievement in English and increased total observed variance from 27% to 32%. The final findings showed that female teachers who rated themselves higher on Instructional Planning and Strategies, and Subject Matter Knowledge, their students also showed higher level of student achievement in English. On the other hand, male teachers who rated themselves at lower levels of Instructional Planning and Strategies, and Subject Matter Knowledge, their students also demonstrated lower level of achievement in English. Gender has always been a variable of interest of the researchers in Pakistan where boys and girls study in separate public schools. Generally, based on the previous board's results, it is strongly perceived that female students perform better than boys on Board's exams (Aziz, 2010). Therefore, it was not unexpected that female teacher would rate themselves higher than male teachers on one or more teacher evaluation variables.

Teacher experience did not show significant relationship with student achievement in English as well as in mathematics. This finding is similar to the findings of the studies conducted by Gallagher (2004) and White (2004) based on Danielson's Framework for teaching (1996). The researcher used teacher experience as a continuous scale and did not make categories because converting a continuous variable into a categorical variable can cause considerable loss of important information in regression analysis (Royston, Altman, & Sauerbrei (2005).

Overall, the results of the study are encouraging. The findings contribute to better understanding of the complex teacher effectiveness construct through individual as well as collective lens. This preliminary study provided evidence of the reliability of the SITE used to measure teacher quality. The SITE helped produce a significant level of variance among teachers' scores on some of the teacher quality indicators. The results also confirmed the

theoretical assertion as a means of finding that effective teachers in Pakistani public schools not only demonstrate higher levels of performance on certain teacher quality indicators, but also they show higher levels of student achievement, at least, in English. Stronge's (2010) work has potential application of the complex teacher evaluation construct measured through a self-assessment tool. The study includes various implications for policy, practice, and future research.

Implications for Policy

The Ministry of Education (2009) Pakistan, in collaboration with the United Nationals Educational Scientific and Cultural Organization (UNESCO), implemented Strengthening Teacher Education in Pakistan (STEP) with the financial support of the United States Agency for International Development (USAID). Under the STEP project, National Professional Standards for teachers were developed in 2008. The purposes of these National Professional Standards were to improve the quality of education by producing world class teachers, and to analyze the factors that contribute to educational quality to impact student learning outcomes (Ministry of Education, 2009).

Selecting some of those factors which have a sound research base, the Self-assessment Instrument for Teacher Evaluation (SITE) was developed to investigate the extent to which these factors contributed to student learning. The initial findings of this study showed (a) positive relationship between teachers' evaluation scores on SITE and student achievement in English, and (b) mixed—positive as well as negative—results between teacher evaluation scores and student achievement in mathematics. The theoretical arguments for this justification suggest that the findings are supportive and similar to the findings of previous research (Gallagher, 2004; Kimball et al., 2004; Milanowski, 2004; White, 2004).

These findings provide initial evidence of the effectiveness of the teacher quality indicators as designed by the Ministry of Education, Pakistan. The provincial as well as the district governments, however, should ensure that all teachers are provided with the complete document of the National Standards because teachers must have deeper understandings of these Professional Standards in relation to their effectiveness and impact on student learning. Since teachers have never self-evaluated before, initially, the policymakers may adopt and introduce the SITE in public schools, and take initial steps toward implementing these Standards.

Implications for Practice

This study provides encouraging results to practitioners. So far, the Performance Evaluation Report (PER) has been used for teacher evaluation and making decisions for teacher promotion. The PER is believed to be fundamentally flawed as it involves various indicators of teacher characteristics which are not necessarily related to teacher effectiveness (Shinkfield & Stufflebeam, 1995). The SITE, on the other hand, provides a new lens to measure teacher quality because SITE is based on research-based teacher quality indicators (Stronge, 2010). The results of this study are encouraging enough for district authorities to consult school administrators and teachers, test the SITE in the public schools, and provide teachers' an alternative to the PER. The teachers, administrators, and district authorities would be able to evaluate teacher performance in a different way and this may help them to make more accurate and valid decisions for teacher promotion and seniority.

Recommendations for Future Research

This quantitative study is a beginning toward a better understanding of the teacher evaluation construct in the Pakistani context. Based on the findings of this study, the following recommendations are being offered for future research:

1. This study was conducted using the simple and multiple regression analyses. Since students and teachers are hierarchically interrelated such as students are nested in classes, and teachers are nested in schools, a further study might be conducted involving the complex statistical analysis such as Hierarchical Linear Modeling (HLM) technique to investigate teacher's effects on student performance.
2. Teacher education and professional qualification were not variables of interest in this study. They can affect student achievement differently (Darling-Hammond, 1990). Future studies related to measuring the relationship between teacher evaluation and student achievement might be conducted involving teacher qualification as variable of interest along with teacher gender and experience.
3. The SITE might be used to evaluate performance of those teachers who teach other subjects such as Physics, Chemistry, Biology, and Social Studies; teachers' scores on the SITE might be used to correlate student achievement in these subjects.
4. The SITE could also be administered to primary and elementary school teachers across varied subject areas.
5. This study involved nonrandom sampling technique and a relatively smaller sample size. A further study might be conducted with a larger sample size for each subject, selected through a random sampling technique such as stratified sampling technique.
6. A representative sample of teachers from urban and rural areas is also suggested for further studies. The results based on location might provide different results (Aziz, 2010).
7. In Pakistani public high schools, depending on the school policy, especially in urban schools with large numbers of students, the students are grouped in grade 9 based on

their previous achievement in grade 8. High achievers, mostly, select science subjects in grade 9 and are grouped in separate section(s), while low achievers opt arts subjects and are grouped separately. So, if the students are grouped in classes based on their previous achievement, the high achievers will, most probably, perform better than low achievers in grade 9 and 10. Future studies might be conducted based on students' grouping in classes; the scores of science students and their teacher's evaluation scores should be correlated separately, while the scores of arts students and their teacher's evaluation scores should be correlated separately.

8. Another basic purpose of developing the SITE was to use this instrument in American schools as an alternative to the principals' ratings which have shown to be lenient, biased, and flawed (Heneman et al., 2006; Kauchak et al., 1985; Milanowski, 2004; Peterson, 2000). Using a random sampling technique with a larger sample size, the SITE might be used in American public schools. The SITE could, perhaps, provide teachers an opportunity to measure their strengths and weakness and to help them grow. The administrators and the policymakers might find a different lens of measuring teacher quality that might be linked to student achievement and growth.
9. Given that the present study used a quantitative design, it would be appropriate to conduct qualitative studies. Teachers of English and mathematics could be interviewed to identify their perspectives about self-assessment as a part of teacher evaluation.

The findings of this study are mixed but encouraging. In this preliminary study, we found some evidence of the potential use of the SITE used for measuring teacher quality. Though the researcher used the SITE for English as well as mathematics teachers, English teachers, perhaps,

found the SITE more interesting and useful than mathematics teachers. This study involved some limitations such as a smaller sample size, non-randomized sampling technique, and use of English language in the SITE instead of Urdu which is the national language of Pakistanis. If the SITE is translated into Urdu and tested over a large number of teachers, especially the mathematics teachers, selected through a randomized technique, the results might be different from the findings of this study. Also, perhaps it is reasonable to think of testing the SITE on the teachers in other subject areas and varied grade levels; it might help identity any causes related to the language proficiency or other problems. Based on the mixed findings and the limitations, the researcher suggests that any generalizations should be made cautiously.

Final Thoughts

Teacher evaluation is a complex phenomenon that comprises multifaceted factors which contribute to student achievement differently in varying contexts under varying levels of teachers' abilities. Various factors such as teacher qualification, teacher gender, and students' previous achievement can be controlled by the researcher; others are uncontrollable and independently affect student achievement. The policymakers strive for making effective policies that can ensure increased output from the teachers and students. The Ministry of Education Pakistan also carved out new professional standards for teachers to ensure that teachers exercise high quality teaching practices that help students increase their achievement.

The researcher conducted this initial study based on these professional standards and developed Self-assessment Instrument for teacher Evaluation (SITE) to measure teacher effectiveness of Pakistani public school teachers. The SITE-based findings are modest but encouraging and they confirm and extend the international research previously done. Measuring teacher quality is a never ending process. If we want to continue our efforts to improve the

quality of education, we must keep abreast with the national and international research on teacher quality indicators and make every effort to improve the teaching and learning process. Indeed, continuous effort to improving educational systems can ensure a country, like Pakistan, to have an honorable place in the comity of nations.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135. doi:10.1086/508733

Adam, F. J. (2011). *Teacher Evaluation Program (TEP). Marzano's art and science of teaching: Teacher evaluation framework*. Retrieved from Florida Department of Education website: http://www.fldoe.org/profdev/pdf/pa/IndianRiver.pdf

Airasian, P. W., & Gullickson, A. (1997). Teacher self-evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practices* (pp. 215-247). Thousand Oaks, CA: Corwin Press.

Airasian, P. W., & Gullickson, A. (2006). Teacher self-evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practices* (2nd ed., pp. 186-211). Thousand Oaks, CA: Corwin Press.

Alexander, E. R., & Wilkins, R. D. (1982). Performance rating validity: The relationship of objective and subjective measures of performance. *Group & Organization Management, 7*(4), 485-496. doi:10.1177/105960118200700410

Almani, A. S. (2002). *A comparative study of effects of in-service training and the performance of secondary school teachers* (Doctoral dissertation). Retrieved from http://prr.hec.gov.pk/Thesis/2739H.pdf

Ashton, P., & Crocker, L. (1987). Systematic study of planned variations: The essential focus of teacher education reform. *Journal of Teacher Education, 38*(3), 2-8. doi:10.1177/002248718703800302

Aziz, M. A. (2010). *Effects of demographic factors and teachers' competencies on the achievement of secondary school students in Punjab* (Doctoral dissertation). Retrieved from http://prr.hec.gov.pk/Thesis/727S.pdf

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment for teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37-65. doi:10.3102/10769986029001037

Barber, L. W. (1990). Self-assessment. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 216-228). Newbury Park, CA: Sage Publications.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133-180. doi:10.3102/0002831209345157

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, *17*(1), 48-62. Retrieved from http://www.isetl.org/ijtlhe/pdf/IJTLHE8.pdf

Betebenner, D. W. (2007). *Estimation of student growth percentiles for the Colorado student assessment program.* Retrieved from Colorado Department of Education website: http://www.cde.state.co.us/cdedocs/Research/PDF/technicalsgppaper_betebenner.pdf

Bibi, S. (2005). *Evaluation study of competencies of secondary school teachers in Punjab* (Doctoral dissertation). Retrieved from http://prr.hec.gov.pk/Thesis/377.pdf

Bill & Melinda Gates Foundation, Measures of Effective Teaching Project. (2010).*Working with teachers to develop fair and reliable measures of effective teaching*. Retrieved from http://www.metproject.org/downloads/met-framing-paper.pdf

Bill & Melinda Gates Foundation, Measures of Effective Teaching Project. (2012). *Gathering feedback for teaching combining high quality observations with student surveys and achievement gains*. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Black, P. J., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7-73. doi:10.1080/0969595980050102

Board of Intermediate and Secondary Education, Lahore. (2009). *Result gazette: Secondary School Certificate (SSC) Examination*. Lahore: Megna Printing Press.

Board of Intermediate and Secondary Education, Lahore. (2010). *Secondary School Certificate (SSC) Examination: Result notification*. Lahore: Megna Printing Press.

Board of Intermediate and Secondary Education, Lahore. (2011). *Secondary School Certificate (SSC) Examination: Result notification*. Lahore: Megna Printing Press.

Bodine, R. (1973). Teacher self-assessment. In E. House, (Ed.), *School evaluation: The politics and process* (pp. 169-173). Berkley, CA: McCutchan.

Bolino, M. C., & Turnley, W. H. (2003). Counternormative impression management, likeability, and performance ratings: The use of intimidation in an organizational setting. *Journal of Organizational Behavior, 24*(2), 237-250. doi:10.1002/job.185

Bolton, D. L. (1973). *Selection and evaluation of teachers*. Berkeley, CA: McCutchan.

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School Journal, 106*(1), 3-20. doi:10.1086/496904

Brink, P. J., & Wood, M. J. (1998). *Advanced design in nursing research* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 329-375). New York, NY: Macmillan.

Buttram, J. L., & Waters, T. (1997). Improving America's schools through standards-based education. *NASSP Bulletin, 81*(590), 1-6. doi:10.1177/019263659708159002

Callender, J. (2004). Value-added student assessment. *Journal of Educational and Behavioral Statistics, 29*(1), 5. doi:10.3102/10769986029001005

Carroll, J. (1981). Faculty self-assessment. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 180-200). Beverly Hills, CA: Sage Publications.

Castetter, W. B. (1971). *The personal function in educational administration*. New York, NY: Macmillan.

Catt, S., Miller, D., & Schallenkamp, K. (2007). Your are the key: Communicate for learning effectiveness. *Education, 127*(3), 369-377.

Centra, J. A. (1972). *Strategies for improving college teaching.* Washington, DC: American Association for Higher Education.

Centra, J. A. (1973). Self-ratings of college teachers: A comparison with student ratings. *Journal of Educational Measurement, 10*(4), 287-295. doi:10.1111/j.1745-3984.1973.tb00806.x

Centra, J. A. (1977). The how and why of evaluating teaching. *New Directions for Higher Education, 1977*(17), 93-106. doi:10.1002/he.36919771709

Charters, W. W., & Waples, D. (1929). *The commonwealth teacher-training study*. Chicago, IL: University of Chicago Press.

Clandinin, D.J. & Connelly, F.M. (1988). Studying teachers' knowledge of classrooms: Collaborative research, ethics, and the negotiation of narrative. *Journal of Educational Thought, 22*(2A), 269-282.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119-142. doi:10.3102/01623737025002119

Coker, H., Medley, D. M., & Soar, R. S. (1980). How valid are expert opinions about effective teaching? *Phi Delta Kappan, 62*(2), 131-134,149.

Cornett-DeVito, M., & Worley, D. W. (2005). A front row seat: A phenomenological investigation of students with learning disabilities. *Communication Education, 54*(4), 312-333. doi:10.1080/03634520500442178

Covino, E., & Iwanicki, E. (1996). Experienced teachers: Their constructs of effective teaching. *Journal of Personal Evaluation in Education, 10*(4), 325-363. doi:10.1007/BF00125499

Cuban, L. (1992). Curriculum stability and change. In P. W. Jackson (Ed.), *Handbook of research on curriculum*, (pp. 216-247). New York, NY: Macmillan.

Cunningham, L. (1997). In the beginning. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure* (pp. 75-80). Thousand Oaks, CA: Corwin Press.

Daley, G., & Kim, L. (2010). *A teacher evaluation system that works.* Retrieved from TAP: The System for Teacher and Student Advancement website: http://www.tapsystem.org/publications/wp_eval.pdf

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (1990). Teaching and knowledge: Policy issues posed by alternate certification for teachers. *Peabody Journal of Education, 67*(3), 123-154. doi**:**10.1080/01619569009538694

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53*(3), 285-328. doi:10.3102/00346543053003285

Dilshad, R. M. (2010). Assessing Quality of Teacher Education: A Student Perspective. *Pakistan Journal of Social Sciences, 30*(1), 85-97. Retrieved from http://www.bzu.edu.pk/PJSS/Vol30No12010/Final_PJSS-30-1-08.pdf

Ellett, C. D., & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness, and school effectiveness: Perspectives from the USA. *Journal of Personnel Evaluation in Education, 17*(1), 101-128. doi:10.1023/A:1025083214622

Emmer, E. T., Evertson, C. M., & Worsham, M. E. (2003). *Classroom management for secondary teachers* (6th ed.). Boston, MA: Allyn and Bacon.

Festinger, L. A. (1954). A theory of social comparison process. *Human Relations, 7*(2), 117-140. doi:10.1177/001872675400700202.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). New York, NY: McGraw-Hill.

Fullan, M. G. (1993). Why teachers must become change agents. *Educational Leadership, 50*(6), 12-17. Retrieved from

http://msit.gsu.edu/charleswang/knowledge/documents/presentation/read5/added%20reading%205.pdf

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education, 79*(4), 79-107. doi**:**10.1207/s15327930pje7904_5

Geer, J.G. (1988). What do open-ended questions measure? *The Public Opinion Quarterly*, 52(3), pp.365-37. Retrieved from http://discoverarchive.vanderbilt.edu/jspui/bitstream/1803/4055/1/What%20Do%20Open-Ended%20Questions%20Measure.pdf

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

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145. doi:10.3102/01623737022002129

Good, T. L., & Brophy, J. E. (1997). *Looking in classrooms* (7th ed.). New York, NY: Longman.

Gronlund, N. E. (2006). *Assessment of student achievement* (8th ed.). Boston, MA: Pearson.

Guskey, T. R. (2002). Does it make a difference? Evaluating professional development. *Educational Leadership, 59*(6), 45-51. Retrieved from http://www.ibhe.org/grants/NCLBProfile/2008/symposium/Guskey_2002_Evaluating_Professional_Development.pdf

Guskey, T. R. (2007). Multiple resources of evidence: An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measurement: Issues and Practice, 26*(1), 19-27. doi:10.1111/j.1745-3992.2007.00085.x

Haney, W., Madaus, G., & Kreitzer, A. (1987). Charms Talismanic: Testing teachers for the improvement of American education. In E. Z. Rothkopf (Ed.), *Review of Research in Education, 14,* 169-238. doi:10.3102/0091732X014001169

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233-251). New York, NY: Springer-Verlag.

Heneman, H. G. III., Milanowski, A., Kimball, S. M., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge-and skill based pay* (RB-45). Retrieved from University of Wisconsin-Madison, Wisconsin Center for Education Research, Consortium for Policy Research in Education website: http://cpre.wceruw.org/publications/rb45.pdf

Heneman, R. L., Greenberger, D. B. & Anonyuo, C. (1989). Attributions and exchanges: The effects of interpersonal factors on the diagnosis of employee performance. *Academy of Management Journal 32*(2), 466–76. doi:10.2307/256371

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406. doi:10.3102/00028312042002371

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5-30). New York, NY: Springer-Verlag.

Hoodbhoy, P. (1998). Out of Pakistan's Education Morass: Possible? How? In P. Hoodbhoy (Ed.), *Education and the state: Fifty years of Pakistan* (1-22). New York, NY: Oxford University Press.

House, E. R. (1973). *School evaluation: The politics and process*. Berkley, CA: McCutchen.

Hudson, S. (2010). *The Effects of Performance-Based Teacher Pay on Student Achievement.* (SIEPR Discussion Paper No. 09-023). Retrieved from Stanford University, Stanford Institute for Economic Policy Research website: http://www.stanford.edu/group/siepr/cgi-bin/siepr/?q=system/files/shared/pubs/papers/09-023_Paper_Hudson.pdf

Hunt, G. H., Wiseman, D. G., & Touzel, T. J. (2009). *Effective teaching: Preparation and Implementation* (4th ed.). Springfield, IL: Charles C Thomas Publishers.

Jacob, B. A., & Lefgren, L. (2005). *What do parents value in education? An empirical investigation of parents' revealed preferences for teachers* (Working Paper No. 11494). Retrieved from the National Bureau of Economic Research website: http://www.nber.org/papers/w11494.pdf?new_window=1

Johnson, B. L. (1997). An organizational analysis of multiple perspectives of effective teaching: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 69-87. doi:10.1023/A:1007951321381

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards*: *How to assess systems for evaluating educators.* Newbury Park, CA: Sage Publications.

Jumani, N. B. (2007). *Study on the competencies of the teachers trained through distance education in Pakistan.* (Unpublished post doctoral research). Retrieved from http://eprints.hec.gov.pk/3517/1/DR_NABI_BUX_JUMANI_post_doc_rport.pdf

Kauchak, D., Peterson, K., & Driscoll, A. (1985). An interview study of teachers attitudes toward teacher evaluation practices. *Journal of Research and Development in Education, 19*(1), 32-37.

Kimball, S. M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education, 16*(4), 241-268. doi:10.1023/A:1021787806189

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54-78. doi:10.1207/s15327930pje7904_4

Kizilbash, H. H. (1998). Teaching teachers to teach. In P. Hoodbhoy (Ed.), *Education and the state: Fifty years of Pakistan* (102-135). New York, NY: Oxford University Press.

Kleinman, G. S. (1966). Assessing teacher effectiveness: The state of the art. *Science Education, 50*(3), 234-238. doi:10.1002/sce.3730500311

Kunter, M., Baumert, J., & Koller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*(5), 494-509. doi:10.1016/j.learninstruc.2007.09.002

Little, J. W. (1993). Teachers' professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis, 15*(2), 129-151. doi:10.3102/01623737015002129

Loughran, J. (2006). Towards a better understanding of science teaching. *Teaching Education, 17*(2), 109-119. doi:10.1080/10476210600680317

Ludtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in

multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120-131. doi:10.1016/j.cedpsych.2008.12.001

Lumpkin, A. (2007). Caring teachers: The key to student learning. *Kappa Delta Pi Record, 43*(4), 158-160. doi:10.1080/00228958.2007.10516474

Marsh, C., & Willis, G. (2007). *Curriculum: Alternative approaches, ongoing issues*. Upper Saddle River, NJ: Merrill Prentice Hall.

Marzano Research Laboratory. (2010). *What works in Oklahoma schools: Phase I report*. Retrieved from http://www.marzanoresearch.com/documents/PhaseI_WWIO.pdf

Marzano Research laboratory. (2011). *The Marzano teacher evaluation model*. Retrieved from http://pages.solution-tree.com/rs/solutiontree/images/MarzanoTeacherEvaluationModel.pdf

Marzano Research Laboratory. (2011). *What works in Oklahoma schools: Phase II report*. Retrieved from http://www.marzanoresearch.com/documents/Phase_II_OK_State_Report.pdf

Marzano, R. J. (2003). *What works in schools: Translating research into action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2007). *The art and science of teaching: A comprehensive framework for effective instruction.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2011). *Research base and validation studies on the Marzano evaluation model*. Retrieved from http://www.marzanoevaluation.com/files/Research_Base_and_Validation_Studies_Marzano_Evaluation_Model.pdf

Marzano, R. J., D. Pickering, D. J., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., Marzano, J. S., & Pickering, D. J. (2003). *Classroom management that works: Research-based strategies for every teacher.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works:* Research-based strategies for increasing student achievement. Alexandria, VA: Association for Supervision and Curriculum Development.

McCaffrey, D. F., & Hamilton, L. S. (2007). *Value-added assessment in practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project*. Retrieved from RAND Corporation website: http://rand.org/pubs/technical_reports/2007/RAND_TR506.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Retrieved from RAND Corporation website: http://www.rand.org/pubs/monographs/2004/RAND_MG158.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Louis, T. A., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 68-101. doi:10.3102/10769986029001067

McGreal, T. L. (1983). *Successful teacher evaluation*. Alexandria, VA: Association for Supervision and Curriculum Development.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 80*(4), 242-247. Retrieved from http://www.jstor.org/stable/pdfplus/40539630.pdf

Mendro, R. L. (1998). Student achievement and school and teacher accountability. *Journal of Personnel Evaluation in Education*, *12*(3), 257-267. doi:10.1023/A:1008019311427

Messick, S. (1989). Validity. In R. L. Linn (3rd Ed.), *Educational measurement* (pp. 13-103). New York, NY: Macmillan.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53. doi:10.1207/s15327930pje7904_3

Milanowski, A. T. (2004). *Relationships among dimension scores of standards-based teacher evaluation systems, and the stability of evaluation score-student achievement relationships over time* (CPRE-UW Working Paper Series TC-04-02). Retrieved from University of Wisconsin-Madison, Wisconsin Center for Education Research, Consortium for Policy Research in Education website: http://cpre.wceruw.org/papers/AERA04Measurement.pdf

Milanowski, A. T., & Heneman, H. G. III. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education*, *15*(3), 193-212. doi:10.1023/A:1012752725765

Ministry of Education. (1998). *National education policy 1998-2010*. Retrieved from http://planipolis.iiep.unesco.org/upload/Pakistan/Pakistan%20Educational%20Policy%20 1998-2010.pdf.

Ministry of Education. (2004). *Education sector reform: Action Plan (2001-02—2005-06).* Retrieved from http://planipolis.iiep.unesco.org/upload/Pakistan/Pakistan%20Education%20Sector%20Reform%202002-2006.pdf

Ministry of Education. (2005). *National education census 2005 Pakistan*. Retrieved from http://www.pbs.gov.pk/content/national-education-census-2005-pakistan.

Ministry of Education. (2009). *National education policy 2009*. Retrieved from http://planipolis.iiep.unesco.org/upload/Pakistan/Pakistan_National_education_policy_2009.pdf

Ministry of Education. (2009). *National professional standards for teachers in Pakistan*. Retrieved from http://unesco.org.pk/education/teachereducation/files/National%20Professional%20Standards%20for%20Teachers.pdf

Monk, D. H., & King, J. A. (1994). Multilevel teacher resource effects in pupil performance in secondary mathematics and science: The case of teacher subject matter preparation. In R. G. Ehrenberg (Ed.), *Choices and consequences: Contemporary policy issues in education* (pp. 29-58). Ithaca, NY: ILR Press.

National Institute for Excellence in Teaching. (2012). *The effectiveness of TAP: Research summary 2012* . Retrieved from TAP: The System for Teacher and Student Achievement website: http://www.tapsystem.org/publications/tap_research_summary_0210.pdf

Odden, A. (2004). Lessons learned about standards-based teacher evaluation systems. *Peabody Journal of Education, 79*(4), 126-137. doi:10.1207/s15327930pje7904_7

Pearson Education. (2008). *Policy Report: Value-added assessment systems*. Retrieved from Pearson Assessment and Information website: http://www.pearsonassessments.com/NR/rdonlyres/2C54ADBA-9959-4F82-82B3-8E65FB822088/0/ValueAdded.pdf. Author.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Popham, J. W. (1971). Performance test of teaching proficiency: Rationale, development and validation. *American Educational Research Journal, 8*(1), 105-117. doi:10.3102/00028312008001105

PVAAS Statewide Team for Pennsylvania Department of Education. (2011). *Pennsylvania Value-Added Assessment System (PVAAS): Guide to PVAAS public reporting*. Retrieved from Pennsylvania Department of Education website: http://www.portal.state.pa.us/portal/server.pt/community/pa_value-added_assessment_system_(pvaas)/8751

Quirk, T. J., Witten, B. J., & Weinberg, S. F. (1973). Review of studies of concurrent and predictive validity of the National Teacher Examinations. *Review of Educational Research, 43*(1), 89-113. doi:10.3102/00346543043001089

Rahman, F., Jumani, N. B., Akhter, Y., Chisthi, S. H., & Ajmal, M. (2011). Relationship between training of teachers and effectiveness teaching. *International Journal of Business and Social Science, 2*(4), 150-160. Retrieved from http://www.ijbssnet.com/journals/Vol._2_No._4%3b_March_2011/18.pdf

Ranne, S. B. (2011). *An assessment of knowledge transition participation in clinical laboratory science* (Unpublished doctoral dissertation). University of Georgia, GA.

Raths, L. (1941). The revised Ohio teaching record. *Educational Research Bulletin, 20*(9), 241-248. Retrieved from www.jstor.org/stable/1474105

Royston, P., Altman, D. G., Sauerbrei, W. (2005). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine, 25*, 127-141. doi:10.1002/sim.2331

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116. doi:10.3102/10769986029001103

Sanders, J. R., & Sullins, C. D. (2005). *Evaluating school programs: An educator's guide*. Thousand Oaks, CA: Corwin Press.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-Model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299-311. doi:10.1007/BF00973726

Sanders, W. L., & Horn, S. P. (1995). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. In A. J. Shrinkfield and D. L. Stufflebeam (Eds.), *Teacher evaluation: A guide to effective practices* (pp. 337-350). Boston. MA: Kluwer Academic Publishers.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement* (Research progress report). Retrieved from the Heartland Institute website: http://heartland.org/sites/all/modules/custom/heartland_migration/files/pdfs/3048.pdf

Sanders, W., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for educational evaluation and

research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256. doi:10.1023/A:1008067210518

Schacter, J., & Thum, Y. M. (2004). Paying for high and low quality teachers. *Economics in Education Review, 23*(4), 411-430. doi:10.1016/j.econedurev.2003.08.002

Schalock, H. D., & Schalock, M. D. (1993). Student learning in teacher evaluation and school improvement: An introduction. *Journal of Personnel Evaluation in Education, 7*(2), 103-104. doi:10.1007/BF00995298

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 244-271). Beverly Hills, CA: Sage Publications.

Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education, 1*(1), 9-23. doi:10.1007/BF00143275

Shinkfield, A. J., & Stufflebeam, D. L. (1995). *Teacher evaluation: Guide to effective practice*. Norwell, MA: Kluwer Academic Publishers.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14. doi:10.3102/0013189X015002004

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 7*(1), 1-22. Retrieved from http://people.ucsc.edu/~ktellez/shulman.pdf

Stodolsky, S. S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher, 13*(9), 11-18. doi:10.3102/0013189X013009011

Stronge, J. H. (2002). *Qualities of effective teachers.* Alexandria, VA: Association for Supervision and Curriculum Development.

Stronge, J. H. (2006). Teacher evaluation and school improvement. In J. H. Stronge (2nd ed.), *Evaluating Teaching: A guide to current thinking and best practice* (pp. 1-22). Thousand Oaks, CA: Corwin Press.

Stronge, J. H. (2007). *Qualities of effective teachers* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Stronge, J. H. (2010). *Evaluating what good teachers do: Eight research-based standards for assessing teacher excellence*. Larchmont, NY: Eye on Education.

Stronge, J. H., & Tucker, P. D. (1995). Performance evaluation of professional support Personnel: A survey of the states. *Journal of Personnel Evaluation in Education*, *9*(2), 123-137. doi:10.1007/BF00972655.

Stronge, J. H., & Tucker, P. D. (2000). *Teacher evaluation and student achievement*. Washington, DC: National Education Association.

Stronge, J. H., & Tucker, P. D. (2003). *Handbook on teacher evaluation: Assessing and improving performance.* Larchmont, NY: Eye on Education.

Stronge, J. H., & Xu, X. (2011). *Teacher keys effectiveness system: Research synthesis of Georgia teacher assessment on performance standards*. Retrieved from Georgia Department of Education website: http://www.doe.k12.ga.us/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/TKES%20Research%20Synthesis_7-11-2012.pdf

Stronge, J. H., Ward, T. J., Tucker, P. D., & Hindman, J. L. (2008). What is the relationship between teacher quality and student achievement? An exploratory study. *Journal of Personnel Evaluation in Education, 20*(3-4), 165-184. doi:10.1007/s11092-008-9053-z

Sykes, G. (1997). On trial: The Dallas value-added accountability system. In J. Millman (Ed.), *Grading Teachers, grading schools: Is student achievement a valid evaluation measure*. (pp. 110-119). Thousand Oaks, CA: Corwin Press.

Tanner, D. (1986). Are reforms like swinging pendulums? In H. J. Walberg and J. W. Keefe, (Eds.), *Rethinking reform: The principal's dilemma* (pp. 5-17). Reston, VA: NASSP.

Teddlie, C., Stringfield, S., & Burdet, J. (2003). International comparison of the relationship among educational effectiveness, evaluation and improvement variables: An overview. *Journal of Personnel Evaluation in Education*, *17*(1), 5-20. doi:10.1023/A:1025020928735

Tomlinson, C. A. (1999). *The differentiated classroom: Responding to the needs of all learners*. Alexandria, VA: Association for Supervision and Curriculum Development.

Tomlinson, C. A. (2007). Learning to love assessment. *Educational Leadership, 65*(4), 8-13.

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. Alexandria, VA: Association of Supervision and Curriculum Development.

Tyack, D. B., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

U.S. Department of Education. (2009). *Race to the Top program executive summary*. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

United Nations Educational Scientific and Cultural Organization. (2006). *Situation analysis of teacher education: Towards a strategic framework for teacher education and professional development*. Retrieved from UNESCO Islamabad website: http://unesco.org.pk/education/documents/step/SituationAnalysis-StrategicFrameworkforTeacherEducation.pdf

Varma, A., & Stroh, L. K. (2001). The impact of same-sex LMX dyads on performance evaluations. *Human Resource Management, 40*(4), 309-320. doi:10.1002/hrm.1021

Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership*, 41(8), 19-27. Retrieved from http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_198405_walberg.pdf

Walberg, H. J., & Paik, S. J. (1997). Assessment requires incentives to add value: A review of the Tennessee Value-added assessment system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure*? (pp. 169-178). Thousand Oaks, CA: Corwin Press.

Walker, L. D., & Avant, K. C. (1983). S*trategies for theory construction in nursing.* Norwalk, CT: Appleton-Century-Crofts.

Waltz, C. F., Strickland, O., & Lenz, E. R. (1984). *Measurement in nursing research.* Philadelphia, PA: F.A. Davis Co.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1994). What helps student learn. *Educational Leadership, 51*(4), 74-79. Retrieved from http://ww.casdk12.net/GHS04/SRB/5-Curriculum/What%20Helps%20Students%20Learn.pdf

Wenglinksy, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10*(12), 1-30. http://epaa.asu.edu/ojs/article/view/291/417

Wenglinsky, H. (2004). The link between instructional practice and the racial gap in middle schools. *Research in Middle Level Education Online*, *28*(1), 1-18. Retrieved from http://www.amle.org/portals/0/pdf/publications/RMLE/rmle_vol28_no1_article1.pdf

White, B. (2004). *The relationship between teacher evaluation scores and student achievement: Evidence from Coventry, RI*. Retrieved from University of Wisconsin-Madison, Consortium for Policy Research in Education website: http://cpre.wceruw.org/papers/CoventryAERA04.pdf

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Berstein, H. T. (1984). *Teacher evaluation: A study of effective practices*. Retrieved from RAND Corporation website: http://www.rand.org/pubs/reports/2006/R3139.pdf

Worley, D., Tistworth, S., Worley, D. W., & Cornett-DeVito, M. (2007). Instructional communication competence: Lessons learned from award-winning teachers. *Communication Studies, 58*(2), 207-222. doi:10.1080/10510970701341170

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teachers and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57-67. doi:10.1023/A:1007999204543

Wright, S. P., Sanders, W. L., & Rivers, J. C. (2006). Measurement of academic growth of individual students toward variable and meaningful academic standards. In R. W. Lissitz (Ed.), *Longitudinal and value-added models of student performance* (pp. 385-389). Maple Grove, MN: Jam Press.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: American Council on Education/Praeger.

Yoon, K. S., Duncan, T., Lee, S. W., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement.*

Washington, DC: Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/rel_2007033.pdf

Zacharias, N. T. (2007). Teacher and student attitudes toward teacher feedback. *RELC Journal, 38*(1), 38-52. doi:10.1177/0033688206076157

Zepeda, S. J. (2006). Classroom Based assessment of teaching and learning. In J. H. Stronge (Eds.), *Evaluating teaching: A guide to current thinking and best practices* (pp. 101-124). Thousand Oaks, CA: Corwin Press.

Zepeda, S. J. (2007). *The principal as instructional leader: A handbook for supervisors* (2nd ed.). Larchmont, NY: Eye on Education.

Zepeda, S. J. (2012). *Professional Development: What works* (2nd ed.). Larchmont, NY: Eye on Education.

Zepeda, S. J. (2012). *The principal as instructional leader: A handbook for supervisors* (3rd ed.). Larchmont, NY: Eye on Education.

APPENDICES

APPENDIX A

SELF-ASSESSMENT INSTRUMENT FOR TEACHER EVALUATION (SITE)

Self-Assessment Instrument for Teacher Evaluation (SITE)

Dear Teacher:

This questionnaire has been designed to understand how frequently you perform the following roles in your school. The information taken from this questionnaire will be kept confidential and no part of this questionnaire will be shared with anyone. Therefore, please read each item carefully and rate your level of performance freely and fairly. **Thank you**.

1. Job Title (e.g. SST)._____2. Qualification: ___3. Subject taught: English □ Mathematics □

4. Class /Section: _____5. Teaching experience this subject (years): ____6. Class Size _____

7. Students appeared in group: Morning □ Evening □ 8. School Name (Code): _____

#	Items	Never	Rarely	Some-times	Often	Always
1	I demonstrate accurate knowledge of my subject matter in lesson plans and teaching.					
2	I demonstrate ability to link present content with past and future learning experiences.					
3	I demonstrate a variety of skills relevant to my subject area(s).					
4	I communicate content in ways that students can understand easily.					
5	I use school and community resources to help students meet their learning needs.					
6	I teach according to the intellectual, emotional, and physical development needs of my students.					
7	I use a variety of teaching strategies to enhance students' understanding.					
8	I change my teaching methodology to make topics relevant to students' lives.					
9	I understand students' individual differences and teach them accordingly.					
10	I use appropriate material, technology, and resources while teaching.					
11	I engage, motivate, and maintain students' attention to their lesson.					
12	I teach the required curriculum according to the time table.					
13	I use maximum time of a period in teaching.					
14	I conduct class tests to monitor student performance regularly.					
15	I evaluate students' performance and provide timely feedback on their errors.					

	Items	Never	Rarely	Some-times	Often	Always
16	I maintain a record of students' results and use it for their future improvement.					
17	I revise content to enhance students' achievement.					
18	My high achieving students evaluate class tests of their class-fellows.					
19	I do not have time to check students' homework.					
20	I keep official record of students' learning progress.					
21	I create a climate of mutual trust and respect in the classroom.					
22	I emphasize continuous improvement toward student achievement.					
23	I maintain a classroom setting that minimizes disruption.					
24	I create a attractive, friendly, and supportive classroom environment.					
25	I ensure students' participation in the learning process.					
26	I encourage students to interact respectfully with each other.					
27	I ensure that lower-achieving students have opportunities to be successful.					
28	I use correct vocabulary and grammar in speaking and writing.					
29	I explain concepts and lesson content in a logical sequence.					
30	I explain lessons according to the age and ability of the students.					
31	I respond to students' questions in appropriate language.					
32	I share students' performance with their parents.					
33	I communicate content problems with my colleagues.					
34	I communicate colleagues in order to improve students' performance.					
35	I look for opportunities for professional growth to enhance my knowledge and teaching skills.					
36	I follow all government policies and requirements.					
37	I maintain confidentiality of all the records of my school.					
38	I respect my community and my school.					
39	I evaluate my strengths and weaknesses and set goals for improvement.					
40	I serve in school committees and support school activities and events.					
41	I participate in decision making committees or groups.					

Comments:

**************Thank you very much **************

APPENDIX B

IRB APPROVAL

Gmail by Google

Muhammad Akram <akramue@gmail.com>

IRB Approval--Zepeda/Akram

9 messages

Megan elizabeth Mcfarland <meganmcf@uga.edu> Tue, Jul 24, 2012 at 9:29 AM
To: "Muhammad Akram (akramue@gmail.com)" <akramue@gmail.com>, MUHAMMAD AKRAM <akram@uga.edu>, "Sally J. Zepeda" <szepeda@uga.edu>

PROJECT NUMBER: 2012-10970-0
TITLE OF STUDY: The relationship between teacher evaluation scores and student achievement: Evidence from Pakistan
PRINCIPAL INVESTIGATOR: Dr. Sally J. Zepeda

Dear Dr. Zepeda and Mr. Akram,

The University of Georgia Institutional Review Board (IRB) has reviewed and approved your above-titled proposal through the exempt (administrative) review procedure authorized by 45 CFR 46.101(b)(2) - Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless (i) the information obtained is recorded in such a manner that human participants can be identified, directly or through identifiers linked to the participants; and (ii) any disclosure of the human participants' responses outside the research could reasonably place the participants at risk of criminal or civil liability or be damaging to the participants' financial standing, employability, or reputation.

Note: Please note the minor revisions made to the consent form and recruitment script; the approved versions are attached. Please be sure these are the versions that you use and that you save them for any future amendment requests. Thanks!

Your approval packet will be sent by mail. Please remember that any changes to this research proposal can only be initiated after review and approval by the IRB (except when necessary to eliminate apparent immediate hazards to the research participant). Any adverse events or unanticipated problems must be reported to the IRB immediately. The principal investigator is also responsible for maintaining all applicable protocol records (regardless of media type) for at least three (3) years after completion of the study (i.e., copy of approved protocol, raw data, amendments, correspondence, and other pertinent documents). You are requested to notify the Human Subjects Office if your study is completed or terminated.

Good luck with your study, and please feel free to contact us if you have any questions. Please use the IRB number and title in all communications regarding this study.

Regards,

Megan

APPENDIX C

AUTHORIZATION LETTER

AUTHORIZATION LETTER

Dear Headmaster/Headmistress

I, Mr. Zafar Masood Anjum, am voluntarily involved in the project of Mr. Muhammad Akram, a Doctoral student in Educational Administration, and Policy, College of Education, University of Georgia, USA. He is going to conduct a research study on the "Relationship between teacher evaluation scores and student achievement: Evidence from Pakistan." On his behalf, I would collect data from teachers who taught English or Mathematics to 10th graders in 2011-12.

If you allow your teachers to participate in this project, they will be asked to complete a Self-assessment Instrument for Teacher Evaluation (SITE) which will take not more than 20 minutes to complete. The achievement scores of 10th graders in English and mathematics will also be taken from these teachers so that students' scores can be linked with the teacher's evaluation scores. There is not any risk attached with the study. All the data will be kept strictly confidential and no part of the data will be shared with anyone. The scanned copies of the data will be emailed to the co-principal investigator (Mr. Akram). If any participant or school personnel are interested in the summary results of this project, they may email Mr. Akram. Upon the completion of the study, Mr. Akram will distribute summary results to the interested personnel to share what he has learned about teacher effectiveness and student achievement.

I will answer any questions about the research now, or during the course of the project, and can be reached by telephone at 0344-6768068 or zmanjum65@gmail.com. You may also contact Mr. Muhammad Akram (akram@uga.edu or +1-706-461-4617) and/ or his professor Dr Sally J. Zepeda (szepeda@uga.edu or 706-542-0408).
Thank you very much

Sincerely

Zafar Masood Anjum
Secondary School Teacher (SST)
Government High School Depalpur
District Okara, Punjab, Pakistan
0344-6768068

I understand the project described above. My questions have been answered to my satisfaction, and I agree to allow my teachers to take part in this study and provide 10th graders' achievement scores in English and math. I have been given a copy of this form to keep.

| --- | --- | --- |
| Name of Headmaster/Headmistress **(Stamp)** | Signature | Date |

APPENDIX D

TEACHER RECRUITMENT LETTER

TEACHER RECRUITMENT LETTER

Hi:

My name is Zafar Masood Anjum, and I am working on a research project with a doctoral student, Mr. Muhammad Akram, and his professor Sally J. Zepeda through the University of Georgia, USA. This project is studying "The relationship between teacher evaluation scores and student achievement: Evidence from Pakistan." I would like to invite you to participate in this research study by completing a 15-20 minutes questionnaire. Moreover, you will be asked to provide students' overall achievement scores in English/Mathematics in grade 10. If you are interested in participating, there is a consent form with further information. I would be happy to answer any questions or you could call me at 0334-6768068 or email at zmanjum65@gmail.com.

Thank you very much.

Regards

Zafar Masood Anjum
Secondary School Teacher
Government High School, Depalpur
Okara

APPENDIX E

VERBAL RECRUITMENT SCRIPT

URDU TRANSLATION

VERBAL RECRUITMENT SCRIPT

URDU TRANSLATION

میرا نام ظفر مسعود انجم ہے اور میں محمد اکرم (پی ایچ ڈی سٹوڈنٹ، یونیورسٹی آف جارجیا، امریکہ) اور ان کی پروفیسر ڈاکٹر Sally J Zepeda کے ساتھ ایک تحقیقی پراجیکٹ میں کام کر رہا ہوں۔ پراجیکٹ کا نام ہے:

The Relationship Between Teacher Evaluation Scores and Student Achievement: Evidence from Pakistan

میں آپ کو اس تحقیقی پراجیکٹ میں شرکت کی دعوت دیتا ہوں۔ اس میں آپ کو ایک سوالنامہ پر کرنا ہوگا جو پندرہ سے بیس منٹ میں مکمل ہو جائے گا۔ اس کے علاوہ آپ سے آپ کی دسویں کلاس کے طلبہ کے انگریزی یا ریاضی کے نمبر بھی لیے جائیں گے۔ اگر آپ اس تحقیقی پراجیکٹ میں شرکت کے لیے دلچسپی رکھتے ہیں تو میرے پاس رضامندی کا ایک فارم بھی ہے جو آپ کو مزید معلومات دے گا۔ میں آپ کے کسی قسم کے سوالات کے جواب دینے میں خوشی محسوس کروں گا۔ اور یا آپ مجھے 0344-6768068 پر کال یا zmanjum65@gmail.com پر ای میل کر سکتے ہیں۔

آپ کا شکریہ

ظفر مسعود انجم (ایس ایس ٹی)

گورنمنٹ ہائی سکول دیپال پور (اوکاڑہ)

APPENDIX F

CONSENT FORM

CONSENT FORM

I, _____, agree to participate in a research study titled " THE RELATIONSHIP BETWEEN TEACHER EVALUATION SCORES AND STUDENT ACHIEVEMENT: EVIDENCE FROM PAKISTAN " conducted by Muhammad Akram, from the Department of Lifelong Education Administration and Policy at the University of Georgia (706-461-4617) under the direction of Dr. Sally J Zepeda, the Department of Lifelong Education Administration and Policy at the University of Georgia (706-542-0408). I understand that my participation is voluntary. I can refuse to participate or stop taking part at anytime without giving any reason, and without penalty or loss of benefits to which I am otherwise entitled. I can ask to have all of the information about me returned to me, removed from the research records, or destroyed.

The reason for this study is to examine the relationship between teacher evaluation scores and student achievement and find out if teacher's evaluation score on the self-assessment questionnaire has relationship with students' achievement in English/Mathematics. If I volunteer to take part in this study, I will be asked to do the following things:

1. Complete the Self-Assessment Teacher Evaluation Questionnaire which will take 15-20 minutes to complete.
2. Provide deidentified student achievement scores of 10th graders whom I taught English/math during the year 2011-12
3. Someone from the study may call me to clarify my information

Once data collection has been completed, all research data including my self-assessment questionnaire will be emailed, through password protected email, to Muhammad Akram. Internet communications are insecure and there is a limit to the confidentiality that can be guaranteed due to the technology itself. However once the materials are received by the researcher, standard confidentiality procedures will be employed.

My score on the self-assessment questionnaire will be used to link with deidentified student achievement scores in English/mathematics. My name and identity will not be shared with my school system or department and researchers will use the data solely for research purposes.

If I am interested in the summary results of this study, I understand that I may email Mr. Akram at akram@uga.edu. The benefits for me are that I would be able to know the strength of the relationship of teacher self-assessment and students' achievement in English/mathematics. The researcher also hopes to learn more about teacher effectiveness and its relationship with students' achievement in English/Mathematics.

No risk is expected and participation in this research will not affect my standing within the school system.

No individually-identifiable information about me, or provided by me during the research, will be shared with others without my written permission. I will be assigned an identifying number and this number will be used on the self-assessment questionnaire.

The investigator will answer any further questions about the research, now or during the course of the project.

I give my permission for the researchers to use scores of self-assessment questionnaire and students' achievement in English/math.
Circle one: YES / NO. Initial _____.

I understand that I am agreeing by my signature on this form to take part in this research project and understand that I will receive a signed copy of this consent form for my records.

_____	_____	_____
Name of Researcher	Signature	Date

Telephone: _____
Email: _____

_____	_____	_____
Name of Participant	Signature	Date

Please sign both copies, keep one and return one to the researcher.

Additional questions or problems regarding your rights as a research participant should be addressed to The Chairperson, Institutional Review Board, University of Georgia, 629 Boyd Graduate Studies Research Center, Athens, Georgia 30602; Telephone (706) 542-3199; E-Mail Address IRB@uga.edu.