Principals' Perceptions of Teacher Ineffectiveness in Elementary Classrooms and How They Relate to Specific Content Areas

by

Steven Franklin

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2014

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree Doctor of Philosophy.

Georgiana Shick Tryon, Ph.D.

Date

Chair of Examining Committee

Alpana Bhattacharya, Ph.D.

Date

Executive Officer

Dr. Jay Verkuilen

Dr. Marian Fish

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

The City University of New York

Abstract

Principals' Perceptions of Teacher Ineffectiveness in Elementary Classrooms and How They Relate to Specific Content Areas

by

Steven Franklin

Advisor: Georgiana Shick Tryon, Ph.D.

The current dissertation was effected to contribute to the existing literature on teacher evaluation. More specifically, the study utilized principals' perceptions to identify what principals, who often evaluate teachers, believe are the most frequent causes of teacher ineffectiveness. For this dissertation, the researcher extended a study by Torff and Sessions (2005). In that study, the authors measured principals' perceptions of the causes of teacher ineffectiveness within high school classrooms. This study extended Torff and Sessions' (2005) research by including elementary school principal perceptions, investigating whether differences exist in elementary school principals' perceptions when asked to rate teacher ineffectiveness across specific academic content areas. Utilizing an ordinal probit model the researcher determined that the only variable that significantly predicted principal perception was Dimension (rating criterion). In addition, the results revealed that, when the researcher controlled for principals' propensity to use the scale in different ways, Implementation Lesson Plans and Writing Lesson Plans were the most frequently rated causes of teacher ineffectiveness across all Domains.

Table of Contents

List of Tables

CHAPTER I

Introduction

Recent trends in educational policy have seen an increased focus on the evaluation of teachers. More specifically, state and national lawmakers have focused much of their attention on ways to identify the most effective teachers. This issue is particularly salient given the current budget crisis. New York State has been a recent battleground for teacher evaluation methods. On March 1st, 2011, the New York State Senate passed legislation to repeal the Last In First Out (LIFO) law in New York City (Majority Press, 2011). This legislation would eliminate the current law, which bases teacher layoffs on seniority alone, and instead institute a merit-based system. This situation has led to a lively discussion, particularly among teachers' unions and state lawmakers highlighting the current trend in educational policy.

The evaluation of teachers has also been a focal point of the federal educational policy discussion. For example, the recent federal Race to the Top program allows a 4.35 billion dollar fund for distribution among state education agencies that institute innovative educational reform. One of the criteria for receiving federal funding as part of this program is for state education agencies to develop methods to evaluate and reward their most effective teachers (U.S. Department of Education, 2009). The criteria for funding are based on a 500-point scoring system, with 138 of those points dedicated to the "Great Teachers and Leaders" criteria (U.S. Department of Education, 2009).

The above are only a few examples of the teacher evaluation debate existing in the popular media today. The identification, and in some cases rewarding, of highly effective teachers has been established as a crucial part of educational reform (Darling-

Hammond, 1999). The next logical step is to determine how to evaluate teachers in the fairest way. Previous research has investigated different approaches to the evaluation of teachers; however, there is not a consensus on which approach is best (Darling-Hammond, 1999). According to Sartain, Stoelinga, and Brown (2011), traditional teacher evaluation systems have done a poor job of differentiating low and high performing teachers. They conducted a study that revealed that 93% of all teachers in the Chicago public schools were in the top two categories of a performance evaluation rating scale with only .3% identified as unsatisfactory. The authors concluded that presently teacher evaluation systems under identify ineffective teachers. In addition, the systems are not able to differentiate various levels of effectiveness, as the vast majority of teachers were categorized in the top two categories, which highlights a potential validity concern. Another study conducted by Markow, Macia, and Lee (2013) found that 98% of principals surveyed gave positive ratings to the classroom teachers in their schools, and 53% of principals identified evaluated teachers as either "challenging" or "very challenging". It is important to provide useful teacher evaluations that identify both effective and ineffective teachers. In the case of ineffective teachers, it is not sufficient to know merely that they are not effective. We need to know the reasons for their ineffectiveness so that programs can be developed to address these issues.

There are inherent flaws in the present approach to teacher evaluations. As discussed, teacher evaluation systems are currently driven by policy makers. As a result, there have been numerous attempts to simplify and streamline these systems. Consequently, policy makers have been unable to come to an agreement, because each system has benefits and inevitable costs. As the following sections detail, these systems

if used alone contain drawbacks that can lead to ill-informed decisions. In other words, the consequential validity of using any one of the various teacher evaluation systems in isolation is poor. This conclusion was first introduced by Donald Campbell, a social scientist who studied program evaluation. The following statement summarizes Campbell's beliefs in relation to evaluating social systems:

> It is in the area of methodological problems generated by political considerations that the assumptions of universality for the methodological principles will fail as we compare experiences from widely differing social, economic, and political systems. (Campbell, 1976, p. 5)

Campbell went on to discuss how political influence affects the validity of social decision-making procedures. This idea is captured by the following statement that is known as Campbell's law:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor. (Campbell, 1976, p. 54)

Campbell argued that psychologists have a unique skill set that allows for fair and unbiased evaluations of social systems. Specifically, he pointed to psychologists' self-criticism as well as caution when interpreting their own measures that has led to the development of mathematical models of reliability and validity, which are important for program evaluation (Campbell, 1976).

Campbell's ideas should resonate with school psychologists who recognize the importance of conducting assessments that use multiple-methods to allow for individual learning differences and environmental causes (Gregory, 2004; Sattler, 2001). In fact,

education law states that psycho-educational assessment used for special education planning cannot be based on any single measure (IDEA; 2004).

In similar fashion, a multifaceted and comprehensive approach needs to be used for teacher evaluations. As Wilkerson, Manatt, Rogers, and Maughan (2000) stated:

> Multisource assessments that tap the collective wisdom of supervisors, peers, students, parents, and others provide the opportunity to more effectively improve teaching and document its quality. Presumably, the best teacher evaluation systems would provide evidence of teaching through documentation of learning – both student learning and teacher learning. Teacher assessments would include not only descriptions of what teachers did in particular episodes over a period of time but also evidence or data of student performance on student outcomes, tying the student performance to actual teaching that preceded the outcomes through a variety of ratings and evaluations of teachers. Multiple feedback sets that include data not only on teacher performance but also the relationship of teacher performance to student performance is not only desirable but also demanded by today's sophisticated parents and interest groups. (p. 181)

If we do not base teacher evaluations on multiple methods, we are placing educators at risk for invalid assessments. In order to evaluate teachers, one first needs to define the criteria on which teachers will be judged. The decision regarding which teacher evaluation methods to use is an area in which there is a great diversity of opinion. The author of the current study promotes the use of a multifaceted system. Historically, however, teacher evaluation systems were often used in isolation. Many of these methods use social psychology constructs such as student achievement and teacher and

administrator perceptions (Darling-Hammond, 1999; Glass, 1990). More recently, however, statistical methods have attempted to replace the social psychology constructs (Chetty, Friedman, & Rockoff, 2011; Kupermintz, 2003). A brief description of teacher evaluation methods appears below.

One evaluation method focuses on students' test scores as indicators of teacher performance. Each state is required to have federally funded schools assess students, through the use of state standardized tests, in grades 3-12 (No Child Left Behind [NCLB], 2003). As a result, all students in these grades are evaluated and teachers are held accountable for student performance. Measuring teacher performance based on students' scores is a logical option and one that can be implemented easily. Darling-Hammond (1999) reported some evidence that supports this system of teacher evaluation. The author reported that partial correlations support strong, significant relationships between teacher quality variables and student achievement. More specifically, the most consistently significant predictors of student achievement (as measured by state-wide standardized tests) in reading and math were the proportions of teachers with full certification and a major in their field of study (*r*s between .61 and .80, $p < .001$) (Darling-Hammond, 1999).

The use of students' test scores to evaluate teachers presents difficulties, however, due to a plethora of confounding variables. Torff and Sessions (2005) identified variables such as students' socioeconomic status and class size that may also play roles in student achievement. The state of Tennessee sponsored a study of class size and student achievement (Achilles & Finn, 2007). In this study, known as The Tennessee Student Teacher Achievement Ratio (STAR), groups of kindergarteners through third graders

were randomly assigned to large classes (22-24 students) or small classes (14-16 students). The study was longitudinal in design and spanned the years 1985-1997. The results found higher test performances of students in the smaller classes, and students assigned to smaller classes maintained their performance advantage throughout their school years.

Darling-Hammond (1999) indicated other factors that relate to student achievement in addition to class size. These include student characteristics such as poverty, non-English language learners, and student minority status. Furthermore, Darling-Hammond (1999) indicated that per pupil spending shows a significant positive relationship with students' scores in reading. These are just a few examples of the many variables that may play a role in student achievement data besides effective teaching. As a result, it is difficult and problematic to use only student achievement data to evaluate teachers.

In an attempt to correct and account for the above variables, a statistical model known as the value-added approach has also been used to evaluate teacher effectiveness (Hanushek & Rivkin, 2010). In a value-added approach, the overall goal is to assess the effectiveness of the teacher by comparing student gains from year to year. The approach attempts to control for confounding variables by utilizing a mixed-model methodology that enables a multivariate and longitudinal analysis of student achievement data (Wright, Horn, & Sanders, 1997). Kupermintz (2003) further defines the model by indicating that the calculation of teacher effectiveness is a process that combines the estimation of the average performance gains in each school system, and the average performance of each teacher's students relative to the system performance.

Overall, one can see how the value-added model attempts to use achievement data

in a more controlled manner by accounting for additional variables. This method, however, has come under criticism, and many factors need to be considered when using value-added approaches to evaluate teachers. Rothstein (2009) argues that value-added models can still produce biased estimates. The author postulates that non-random assignment of students to teachers can have an effect. For example, if students who experience high achievement gain are assigned to a particular teacher, the following year the estimated value added for the teachers with high prior year gains will be biased downward (Rothstein, 2009). Furthermore, Kupermintz (2003) cites psychometric weaknesses with the value-added model. One major drawback is that value-added calculations are only possible when similar tests from the previous year are available for comparison. If the tests focus on a different skill or content area from one year to the next, the value-added approach loses validity.

Schochet and Chiang (2004) attempted to demonstrate these weaknesses statistically. The authors addressed error rates for measuring teacher and school performance when using value-added models. They developed error rate formulas and ran simulations to assess the likely error rates that would be associated with utilizing a value-added model. The authors found that Type I and Type II error rates for comparing a teacher's performance to the average are likely to be about 25% with three years of data and 35% with one year of data. Despite obvious advances in statistical analyses, the literature still suggests caution when using these methods to make decisions that have lasting consequences.

The final method discussed here and used by the current study is principal evaluation of teachers. Principal evaluation has been used historically; however, many

critics claim that the subjective nature of principal evaluations affects their validity (Darling-Hammond, 1999). In addition, teacher evaluations by principals have historically consisted of inaccurate ratings, which are not context specific (Stodolsky, 1984). Torff and Sessions (2005), however, indicate many benefits of principal evaluation. For example, principals observe teachers' classroom performances. This gives them a first-hand account of vital criteria of teacher effectiveness such as lesson implementation and classroom management. Principals also have access to standardized performances of students in each teacher's classroom. In addition, according to Darling-Hammond and Snyder (2000), principals receive feedback about teacher performance through constant contact with students, parents, other administrators, and department heads. Finally, principals have background experience and training in education. This includes principals' experience as classroom teachers themselves that further informs their perceptions and ratings (Torff & Sessions, 2005).

Current systems that use a standards-based approach have found a correlation between principal ratings of teacher performance and student achievement. Milanowski (2004) provided evidence demonstrating that when principal ratings are based on a comprehensive rubric as part of a larger evaluation system that includes other elements, teacher evaluation scores can be used as predictors of student achievement in math, reading, and science.

The current study sought to add to the literature on teacher evaluation by principals by identifying specific criteria associated with ineffective teaching by elementary school teachers. This study is a replication and extension of a study by Torff and Sessions (2005 & 2009) that identified principals' evaluations of high school

teachers' ineffectiveness. The hope is that identification of specific criteria will lead to the development of more focused and fair principal evaluations of teacher performance within the context of a multifaceted approach. In addition, teacher-training programs may be able to use this information to remediate shortcomings in future teacher candidates. The dissertation study sought to answer the following research questions:

1. According to elementary school principals, what is the most frequently perceived cause of general teacher ineffectiveness?
2. According to elementary school principals, what is the most frequently perceived cause of teacher infectiveness within specific instructional content areas?
3. According to elementary school principals do perceived causes of teacher ineffectiveness differ between instructional content areas (ELA and Math)?

CHAPTER II

Literature Review

The current literature review first outlines the importance of effective teacher evaluation systems by reviewing literature that relates rating systems of teacher effectiveness to student achievement. Next, the chapter reviews different teacher evaluation systems. The review addresses research that describes evaluation systems as well as research that details the pros and cons of each system. The purpose of this section is to highlight the need for a multi-faceted evaluation system, as each of these systems used alone would produce an invalid assessment of teacher performance. Next, the chapter provides a detailed review of a study by Torff and Sessions (2005) that provided the framework for this dissertation as well as for the dissertation pilot study. The chapter concludes with the rationale and hypotheses for the current study.

Why Research Teacher Effectiveness?

As discussed previously, the evaluation of teachers has become a federal education issue. This is exemplified by the federal Race to the Top program, which allows state education agencies that institute innovative educational reform to receive a portion of a 4.35 billion dollar fund. One of the criteria for receiving federal funding as part of this program is for state education agencies to develop methods that evaluate and reward their most effective teachers (U.S. Department of Education, 2009). This law and its ties to federal funding have resulted in an immediate need for more research that addresses the effectiveness of teacher evaluation systems, as well as the criteria that should be used to determine the effectiveness of a teacher. The lack of a consensus to determine what is both effective and fair has prevented some schools from receiving

federal and state funds. For example, in New York City disagreement between the United Federation of Teachers and the Department of Education on how to meet the Race to the Top criteria put 60 million dollars in federal grant money at risk (Santos, 2011). As a result, the money, which was to aid 33 struggling schools across New York City, never reached those schools. Dr. King, the New York State education commissioner, was forced to suspend the grants until the parties could agree upon an evaluation system (Santos, 2011). Similarly, the governor of New York, Andrew Cuomo, decided to withhold an additional 4% increase in state education funding if districts could not agree upon teacher evaluation systems with their unions by 2013 (Fertig, 2012). It was not until a binding arbitration led by Commissioner King, following the failure of the New York City Department of Education to reach an agreement with the United Federation of Teachers, that an evaluation plan was put into place (McAdoo, 2013). The plan was unveiled June 1, 2013 and is discussed later in the literature review.

Another example of the need for teacher evaluation research comes from implementation of new pilot programs by different states. New Jersey is one state with a pilot program. According to the New Jersey Department of Education (2011), 11 districts qualified for participation in the pilot program, named Excellent Educators for New Jersey, that allows qualified districts to split 1.1 million dollars in grant funds. The funds were allocated to districts to develop new teacher evaluation systems that meet a general framework proposed by the State of New Jersey. This framework includes the following general criteria:

- Teachers should never be evaluated on the basis of a single consideration, such as test scores much less a single test, but on the basis of multiple measures that

include both learning outcomes and effective practices, with approximately 50% associated with each.

- Where applicable, the component of the evaluation based on "learning outcomes" should include, but is not limited to, progress on objective assessments such as NJ ASK. In untested grades and subjects, for example, student achievement might include a focus on student work or locally determined criteria.
- To avoid penalizing teachers who work with our highest need students, evaluation criteria should favor student progress and not absolute performance.
- To give teachers meaningful information to help them develop, the prior system of binary ratings (either "satisfactory" or "unsatisfactory") will be replaced by a four-tiered system, including "ineffective," "partially effective," "effective" and "highly effective."
- Districts should provide a direct link between the results of the evaluation and professional development opportunities to help teachers at all levels continuously improve.
- To assure consistency and fairness, plans should address inter-rater reliability – solving for the problem of differences in how individual evaluators review teachers across schools and districts.
- Any personnel consequences connected with evaluations remain a matter of local decision and applicable state law and are not an element of the pilot program. (New Jersey Department of Education, 2011, p. 1).

New Jersey is just one example of a state taking definitive steps to implement a more systematic teacher evaluation program. New Jersey provides a perfect example of

how research on teacher evaluation is both necessary and timely. Despite the State's steps to define what they refer to as a "framework", state education officials still continue to "encourage districts to innovate within that framework" (New Jersey Department of Education, 2011, p. 1).

Teacher Evaluation Methods

In addition, the commitment to implement teacher evaluation programs is evident from the large amount of funds allocated by both individual states and the federal government to their development. Thus, we will now examine various teacher evaluation methods.

Student achievement scores alone. One proposed method to evaluate teacher effectiveness is to use student achievement scores exclusively as indicators. Because it is important for schools to produce students who score well on achievement tests, some believe that using student achievement data is the most effective method to evaluate teachers (Darling-Hammond, 1999; Haertel, 1986). The idea behind this method is that raising student achievement scores is the goal of any school, and therefore teachers should be evaluated based on their ability to influence these scores (Haertel, 1986; Torff & Sessions, 2009). In addition, to increase objectivity, class-level grades are often not used as measures of student achievement. Instead, statewide-standardized achievement data are used. Another reason for using standardized achievement data is that these are the tests scores on which states and schools themselves are evaluated (Haladyna, Nolen, & Hass, 1991).

Objectivity is a crucial benefit of using standardized achievement data alone to evaluate students' knowledge. Popham (1999) indicates that standardized achievement

tests are norm-referenced and so supply a statistically valid reflection of achievement. In addition, the tests provide evidence of students' relative strengths and weaknesses and allow for comparisons among students nationwide. One can use these assessments to measure progress over time. Due to their norm-referenced nature, the assessments can be given from year to year, allowing for comparisons of standardized scores in a teacher's classroom over time.

Criticisms of using student achievement alone. Overall, however, Popham (1999) argues that the use of students' scores on standardized assessments as the sole method to evaluate teacher quality is a generally poor practice. One argument against this practice is that it is impossible for standardized achievement tests to incorporate and align state and local education agency objectives into their assessment. This can result in a mismatch between class-level objectives and assessment objectives. In addition, according to Baker et al. (2010) reliance on test scores leads to the narrowing and oversimplifying of the curriculum and to eventual misidentifying of both successful and unsuccessful teachers.

A second reason for not evaluating teachers solely on standardized achievement test performance is that standardized assessments make general judgments about student achievement from a small sampling of content (Baker et al., 2010; Popham, 1999). Standardized achievement tests are generally formatted to cover a large content domain with a relatively low number of items. Typically, these tests are designed for the student to finish in one test-taking session. In addition, they tend to assess basic skills and complex outcomes adaptable to many local situations and do not reflect timelines of the local curriculum (Miller, Linn, & Gronlund, 2009). As a result, standardized assessments

may not align with the content that the teacher has covered in class.

A third reason for not using standardized achievement data alone to evaluate teacher effectiveness is that other factors also affect student performance. Popham (1999) indicates that, in addition to teacher quality, the curriculum of the particular school, each student's intellectual ability, and students' out of school learning influence achievement. Other factors such as measurement error and instability, the nonrandom sorting of teachers across schools and of students to teachers, as well as the difficulty in understanding the contributions of multiple teachers over time have also lead to skepticism over the use of test scores as the only indicator of teacher performance (Baker et al., 2010).

Expanding Popham's (1999) ideas, Darling-Hammond (1999) investigated other factors that contribute to student achievement in addition to teacher quality. This study used data from a 50-state policy survey conducted by the National Commission on Teaching and America's Future, case studies conducted by the Center for the Study of Teaching and Policy, the 1993-1994 Schools and Staffing Survey, and the National Assessment of Educational Progress. Using these data, Darling-Hammond (1999) conducted bi-variate correlations of school resource variables and student demographic variables with state average test scores. The study revealed that student characteristics such as poverty, non-English language status, and minority status are significantly negatively correlated with student outcome. In addition, Darling-Hammond (1999) found class-size to be another contributing factor to student achievement. Smaller class sizes were moderately, positively correlated with achievement, with the highest effects being in fourth grade reading.

In an attempt to promote a more multi-faceted approach, Haertel (1986) questioned the validity of using standardized achievement scores to evaluate teachers. Haertel's approach to teacher evaluation includes the use of student portfolios, student attendance, and the formation of fair comparison groups. Haertel (1986) first argued that using student achievement gains alone raises the question of how to distinguish achievement due to instruction versus achievement as a result of other factors such as homework completion, self-directed study, and tutoring. The second validity concern that Haertel (1986) presented is the issue of initial student competence. Teachers may need to alter their teaching to provide instruction at the students' instructional levels. Unfortunately, this may result in a deviation from the expected grade level curriculum that results in a mismatch between what is taught and what is on the standardized test. So, though the teacher may engage in the best teaching practice (i.e., teaching to the students' instructional level), the quality of instruction will not be represented by the students' performances on the test.

Haertel (1986) also discussed home support as another validity threat to the sole use of achievement test scores to assess teacher competence. Parental support and expectations may influence student motivation to study, and some parents also provide direct instruction to their children. Typical examples of parental support include reading with children or aiding in homework completion. Other home support examples that affect achievement include, but are not limited to, number of books in the home, provision of a regular place to study, and living with both parents (Haertel, 1986).

Finally, Baker et al. (2010) and Haertel (1986) discussed additional school-wide variables such as quality of instructional materials, time available for instruction, school-

wide learning climate, and instructional support from other teachers that contribute to student achievement. These factors, in addition to those identified by Darling-Hammond (1999), speak to the validity concerns of using student achievement data alone as measures of teacher competence (Baker et al., 2010). The existence of a several confounding variables makes it difficult to isolate the amount of variation in student achievement scores that is a direct result of teacher instruction.

Haladyna, Nolen, and Hass (1991) addressed concerns about using standardized test scores from a different angle. The authors focused on potential negative consequences as a validity threat. Haladyna et al. (1991) argued that the use of standardized test scores to evaluate teachers and schools increases the pressure to raise the test scores, and as a result leads to a phenomenon known as test score pollution. Test score pollution is defined as "...factors affecting the truthfulness of a test score interpretation. Specifically, pollution increases or decreases test performance without connection to the construct represented by the test, producing construct-irrelevant test score variance" (Haladyna et al., 1991, p. 4).

Haladyna et al. (1991) presented a number of polluting practices that take place within schools. Some of these practices are agreed upon as ethical, however, many of them are not. Ethical forms include practices such as training in testwiseness skills (includes familiarizing students with answer sheet formats and teaching general test taking strategies), checking answer sheets to make sure that each has been properly completed, and increasing student motivation to perform on the test through various methods.

The following practices are unethical; however, many of these do not seem

inherently wrong. Here, unethical is defined as those practices that inflate test scores without concurrently raising student achievement. Such practices, according to Haladyna et al. (1991), include teaching test-taking skills, developing a curriculum to match the test, preparing teaching objectives to match the test, presenting items similar to those presented on the test, using commercial materials specifically designed to improve test performance, and presenting before the test the actual items to be tested. Although these practices increase students' test scores, they decrease teacher quality and therefore negatively affect student achievement on all other measures except for the high-stakes, standardized assessment. The authors discussed how teacher quality is affected when schools align their curricula to match the tests. This limits the material and method by which teachers instruct their students. Teachers often end up teaching to the test, while sacrificing individual student needs as well as teacher autonomy in terms of lesson planning and time spent on various content domains (Baker at al., 2010; Haladyna et al., 1991).

Value-added models. Value-added methods of evaluating teacher effectiveness use student test score data, as discussed above, to calculate how much value the teacher adds to student performance outcomes (Chetney, Friedman, & Rockoff, 2011; Hanushek & Rivkin, 2010). Value-added models are statistical methods that attempt to account for variables that may explain student test score gains, as well as factors that may prevent gains, to determine the exact impact of the teacher (Kupermintz, 2003).

Hanushek and Rivkin (2010) provide a brief summarization of value-added systems and how they work. According to the authors, value-added assessments typically use the following base function:

$$A_g = \theta A_{g-1} + \tau_j + S\varphi + X_\gamma + \varepsilon.$$

Here the A_g represents the achievement of student i in grade g; A_{g-1} represents the prior year student achievement in grade g -1; S is a vector representing school and peer factors; X is vector representing family and neighborhood factors; ε is a term representing unmeasured variables; and θ, φ, and γ are unknown parameters. Finally, the statistic that represents a teacher's fixed effect, which provides a teacher value-added measure is τ_j. It is this value that the value-added literature reports, and it is typically expressed in standard deviation units (Hanushek & Rivkin, 2010). According to Hanushek and Rivkin (2010), value-added estimates from a range of school environments across the United States are fairly similar. The average standard deviation for reading is .13 (range = .08 - .26) and for math is .17 (.11 - .36).

These gains appear to provide evidence for the use of value-added methods, however, there is active debate on the usefulness and validity of the approach (Chetney et al. 2011; Hanushek & Rivkin, 2010; Kupermintz, 2003). Chetney et al. (2011) investigated the usefulness of the value added approach by answering the following two research questions: (a) Do value-added estimates provide an unbiased estimate of teachers' impact on student achievement gains? and (b) Do high value-added teachers improve students' long-term outcomes? Chetney et al. (2011) addressed these questions by analyzing data from two large administrative databases. The first database covered classroom assignments from a large urban school district containing data for 2.5 million students. The second data set was U.S. tax records, which contain information on individual earnings, college attendance, and teenage births. In addition, the authors collected parent characteristics such as household income, retirement savings, and

mother's age at child's birth. The authors then matched the school district records to about 90% of the tax data, allowing them to track a large group of students from elementary school to adulthood.

In order to estimate bias, the authors used parent characteristics from the tax data, which have been demonstrated to be strong predictors of achievement, but are not included in value-added models (Chetney et al., 2011). The authors found that parent characteristics were uncorrelated with teacher-value added effects (Chetney et al., 2011). The results showed that a one standard deviation increase in teacher value-added scores only predicted scores based on parent characteristics by .01 standard deviation units. This finding provides evidence that value-added models are not biased. For example, the results demonstrate that two students in the same grade, who have the same test scores, ethnicity, and discipline record do not have systematically different parental characteristics, even if one of those students is assigned to a teacher with a higher value-added score (Chetney et al., 2011).

Chetney et al. (2011) investigated whether value-added scores related to student future outcomes by regressing outcomes, such as earned income, on teacher value-added scores for a given grade. Results indicated that a 1 standard deviation increase in teacher value-added increased the probability of student college attendance by age 20 by .5 percentage points. In addition, another finding revealed that at age 28, a 1 standard deviation increase in teacher quality, indicated by value-added scores, increased annual earnings by 1% on average (Chetney et al., 2011).

Criticisms of value-added methods. Chetney et al. (2011) presented potent findings for the use of value-added methods as one way to evaluate teachers. Kupermintz (2003)

discussed potential limitations of using value added as the only measure of teacher effectiveness. Kupermintz (2003) indicated that student aptitude may be one factor that affects value-added scores. The argument is that students vary on their readiness or ability to respond to instruction. Students vary on cognitive, motivational, and affective factors. The value-added model attempts to control for this by using the individual student's prior year achievement as a measure of aptitude. Thus, the child serves as his or her own control. Kupermintz (2003) indicated that this is a valid approach within an experimental context. Problems arise, however, when certain assumptions are not met within the classroom and school context. These conditions include random assignment of students to teachers or a careful and systematically balanced allocation of students to teachers. Kupermintz (2003) indicated that this is often not the case, which then poses a threat to the validity of the value-added measure. This point is reiterated by Baker et al. (2010) who indicated that, even when student demographics are taken into account, value-added models are too unstable across time classes and across tests to use to evaluate teachers.

According to Kupermintz (2003), another factor that may threaten the accuracy of teacher effects is the amount of available data. Baker et al. (2010) demonstrated that value-added models may be better suited to handle larger aggregations of data on a school or district level. Classroom level data, however, consists of sample sizes that are and can lead to dramatic year-to-year fluctuations and as a result produce unstable estimates of teacher effects. Kupermintz (2003) argued that teacher effects are pulled toward the school system's average. As a result, regression of teacher effects towards the mean may produce unacceptable rates of false-positive and false-negative classifications,

because teacher effects would be more influenced by student population variables such as the class size than by teaching performance. The following study by Schochet and Chiang (2010) attempted to identify error rates when using value-added models.

Schochet and Chiang (2004) attempted to demonstrate these weaknesses statistically. The authors addressed error rates for measuring teacher and school performance when using value-added models. They defined a false positive error rate as the probability of a teacher (whose true value is q standard deviations above average) is falsely being identified as needing special assistance. Schochet and Chiang (2004) defined this using the following formula:

$$FPR(q) = \text{PR}(\text{Reject } H_o \mid \tau_{jk} - = _{\tau} = q\sigma) = 1 - \Phi\,[\Phi^{-1}(1-\alpha),$$

$$+ q\sigma\lambda / \sqrt{v_2}\,] \text{ for } q \leq 0 \text{ (p. 13).}$$

In addition, the authors defined the false negative error rate as the probability that the hypothesis test will fail to identify teachers whose true performances are at least T standard deviations below average. This is the probability that a low performing teacher will not be identified for special assistance when needed (Schochet & Chiang, 2004, p. 13). Schochet and Chiang (2004) represented the false negative rate using the following formula:

$$FPR(q) = \text{PR}(\text{Do Not Reject } H_o \mid \tau_{jk} - = _{\tau} = q\sigma) = 1 - \Phi,$$

$$[\Phi^{-1}(1-\alpha) + q\sigma\lambda / \sqrt{v_2}\,] \text{ for } q \leq T < 0 \text{ (p. 13).}$$

Using these formulas, the authors ran simulations to assess the likely error rates that would be associated with utilizing a value-added model. They found that Type I (false positive) and Type II (false negative) error rates for comparing a teacher's performance to the average are likely to be about 25% with three years of data and 35%

with one year of data.

Baker et al. (2010) further warned that using value-added models may lead to unintended negative effects. As demonstrated above, the amount of data collected can influence the validity of value-added models. The use of value-added models within the context of schools may also introduce a myriad of confounding variables that affect validity. Baker et al. (2010) described how the use of test scores within value-added models evaluate teachers who are working with high needs students unfairly. The inability of value-added models to account for differences in student characteristics and summer learning loss could lead to the inappropriate dismissal of teachers of low-income and special education students.

Principal ratings of teachers. The final method that has been historically used to evaluate teachers is principal ratings. Despite changes in policy and method, principal ratings have remained a stable component of teacher evaluation. In 2000, Wilkerson, Manatt, Rogers, and Maughan indicated that in 46 states only a principal or other single administrator assessed the performance of teachers.

A review of the literature (Danielson & McGreal, 2000; Peterson, 2000, 2004; Stodolsky, 1990) reveals that the way in which principals rate their teachers varies from school to school. In addition, principal ratings have appeared to evolve over time. Stodolsky (1990) described an early method known as "open systems" that was primarily utilized in the 1950s and 1960s. In open systems, principals attempted to record all behaviors that teachers exhibited in any given observational period. These open records were taken in an attempt to describe the teacher's actions fully and were later used to formulate an evaluation or assessment of that teacher (Stodolsky, 1990). In contrast to

the open evaluation systems are what Stodolsky refers to as "closed systems" that were developed more recently. These teacher evaluation systems have many forms including behavior checklists and rating scales. They differ from open systems in that the observer (i.e., principal) only rates certain pre-determined characteristics instead of recording every teacher behavior observed.

Closed systems are frequently used today in the form of a checklist, rating scale, or anecdotal report recorded by an administrator after one or more classroom visits (Peterson, 2004). In fact, an administrator classroom visit with a checklist reporting system is the most common practice of teacher evaluation (Peterson, 2000). Classroom walk-throughs are often utilized as a supplement to the standard checklist systems (Peterson, 2000). A typical walk-through consists of a brief, unscheduled, informal classroom visit lasting 3-6 minutes. Walk-throughs are less intrusive and are able to be carried out on a more frequent basis compared to more formal classroom teacher observations (Peterson, 2000).

According to Danielson and McGreal (2000), many of the methods that principals use today were developed in the 1970s to early 1980s and are based on recording a small number of observable teacher behaviors. This was in response to the pressure to apply a clinical supervision framework to teacher evaluation (Danielson & McGreal, 2000). Policymakers hoped to identify specific teacher behaviors that were linked to student achievement. Research on the relationship between teacher behaviors and student achievement was known as "teacher effects" research and was led by Madeline Hunter, who developed a behaviorist learning theory emphasizing teacher-centered, structured classrooms (Danielson & McGreal, 2000). Unfortunately, many checklists and rating

scales that state and local education agencies developed used only a sampling of teacher behaviors from the teacher effects research, and this selective sampling resulted in an oversimplification of the teaching process and an inability of principals to make accurate assessments of teacher performance (Danielson & McGreal, 2000; Medley & Coker, 1987; Mianowski, 2004; Stodolsky, 1984). Critics suggest that teacher evaluations based solely on a discrete list of behaviors ignore many other factors. These factors include, but are not limited to, student demographic information, student progress over the course of the year, and the teacher's ability to differentiate instruction (Danielson & McGreal, 2000; Medley & Coker, 1987; Mianowski, 2004; Stodolsky, 1984).

This point is exemplified in a recent survey of teachers and principals conducted for Met Life, Inc. (Markow, Macia, & Lee, 2013). The survey was conducted between October 5 and November 11, 2012 via telephone interviews with 1,000 U.S. K-12 public school teachers and 500 K-12 principals (Markow et al., 2013). Results of the survey indicated that demographic variables do play a large role in influencing teacher evaluation scores. Markow et al. (2013) found that principals in schools with more than two-thirds low income students are less likely to give teachers an excellent rating when compared to principals from schools with one-third or fewer low income students (51% vs. 75% of principals surveyed). In addition, Markow et al. (2013) reports that 50% of principals surveyed from schools with a large minority student body are less likely to assign excellent ratings when compared to 72% of principals from schools whose student body is not comprised of a predominantly minority student body.

Medley and Coker (1987) investigated the accuracy of principal ratings as predictors of teacher effectiveness. The variables included in the study were principals'

judgments of teacher performance and direct measures of teachers' effects on students based on achievement test scores in reading and math (Medley & Coker, 1987). Pre- and post-test scores were used to measure achievement on the two standardized achievement tests. Principals' judgments of teacher performance were recorded utilizing a form developed by the authors. The questionnaire asked principals to assign a score from 1 – 20 on three separate teacher roles for each teacher: Role 1 represented the teacher's responsibility for providing learning experiences that led to acquisition of knowledge; Role 2 represented the teacher's responsibility for providing learning experiences that led to good citizenship; and Role 3 represented the teacher as a professional colleague. The score represented where the principal believed the teacher ranked in a representative group of teachers. A score of 1 was equivalent to an inferior performance relative to all other teachers and a score of 20 indicated a superior performance relative to all other teachers (Medley & Coker, 1987). .

Utilizing a sample of 322 teachers, Medley and Coker (1987) found that principals' judgments of teacher performance showed little relationship to students' test scores. More specifically, the authors found a mean correlation of .17 over all roles and students' scores, and a mean correlation of principals' judgments of Role 1 (the ability to promote the acquisition of knowledge) with students' test scores of .20 (Medley & Coker, 1987).

An earlier study by Stodolsky (1984) may provide some explanation of why and how principal ratings are historically inaccurate relative to student achievement. Stodolsky (1984) argued that rating systems assume that teaching skills are the same across various content areas. On the contrary, Stodolsky (1984) suggested that teacher performance is context specific, and based on the subject taught and the students who are

taught. Stodolsky (1984) attempted to study and identify systematic variation in teachers' instructional practice across what she referred to as "activity segments". Stodolsky (1984) defined activity segments as parts of the instructional day that have a specific instructional format, participants, materials, behavior expectations and goals, and space-time boundaries (p. 14). Stodolsky (1984) studied a group of 20 fifth grade classes in math and 19 fifth grade classes in social studies in a Chicago area school district. Trained observers conducted two sets of observations. One set included observations of the activity structure of the classroom, which included use of materials, teacher location, pacing of the lesson, duration of various activities, and student behavior and location. The second set of observations included a strict time-sampling rotation method to observe a subset of children in each classroom to obtain information regarding student behavior and task involvement.

The author described the results in terms of instructional format, pacing, and cognitive level. In general, Stodolsky (1984) found differences in all three categories using the same teacher as the basis of comparison across content areas. For example, observations of the same teacher were compared across different content areas. Within instructional format in math, independent seat-work encompassed 40% of the time while in social studies, independent seat-work only accounted for 18% of the instructional time. Pacing was more student-centered and cooperative in social studies compared to math, which involved more group work. Finally, the author found differences in the cognitive level of segments. In math, close to 80% of the segments had learning skills, concepts, and algorithms as goals (Stodolsky, 1984). In contrast, social studies lessons were more varied with about one third being factual and one quarter involving research skills and

higher-level processes such as evaluation (Stodolsky, 1984). Despite observing different teaching practices, the authors found almost identical levels of student involvement, assessed by on-task behavior, in the two subjects of math (mean percent on task = .77, *SD* = .07) and social studies (mean percent on task = .75, *SD* = .09).

Stodolsky's (1984) results suggest that a major cause of the low predictability of student performance by principals' teacher ratings may be the different expectations for both teachers and students across content areas. As Stodolsky (1984) stated:

> If the teachers observed in our study had been evaluated through an observation of a math lesson or of a social studies lesson, highly discrepant information would have been obtained. A characterization of their teaching methods and skills would not have produced a valid picture that would generalize across the two subjects (or others) they teach. (p. 17)

Teaching variations across contexts have led to a push to alter principal ratings to include observations across content areas. A new method of teacher evaluation with principal ratings as a central component is 'standards-based evaluation' (Danielson & McGreal, 2000). The central component of this evaluation system is a comprehensive rubric rating scale that includes explicit teaching standards. This rubric contains 17 performance standards grouped into four domains: planning and preparation, creating an environment for learning, teaching for learning, and professionalism (Danielson & McGreal, 2000). The rubric spans four levels of performance (*Unsatisfactory, Gaining Proficiency, Proficient, and Exceeds Proficiency*), and is utilized in six separate classroom observations per teacher (Danielson & McGreal, 2000). In addition to multiple classroom observations, the method developed by Danielson and McGreal

(2000) includes the use of teaching artifacts, such as classroom lesson plans and samples of student work, often in the form of student portfolios. Other information utilized to ensure a more comprehensive evaluation includes a teacher self-assessment and student/ parent/colleague feedback.

In contrast to results of previous studies that used principals' evaluations of teachers, recent studies have found promising results using the standards-based evaluation system. Mialnowski (2004) investigated the relationship between science, reading, and mathematics achievement scores on state and district wide tests in grades 3 through 8 and standards based teacher evaluation scores of 212 teachers from the Cincinnati Public Schools in 2000-01 and 2001-02. Milanowski (2004) found the following average positive correlations between teacher evaluation scores and student achievement: .27 for science, .32 for reading, and .43 for mathematics. The teacher assessment system also identified teachers who had students with higher than expected levels of achievement to a degree greater than chance. In sum, the Milanwoski (2004) results indicate that, despite issues in the past, principal evaluation scores can be utilized within a comprehensive and multi-faceted system to accurately represent teacher practices that affect student learning. The following section outlines further studies that link standards-based teacher evaluation scores to student achievement and support the use of valid teacher evaluations as meaningful predictors of student achievement outcomes.

Teacher Evaluation and Student Achievement

All practices that are implemented within the school setting should, to some degree, lead to gains in student achievement. The implementation of teacher evaluation systems is meant to affect student achievement by ensuring that schools have the highest quality

teachers through the assessment of teacher quality and the provision of feedback to teachers to help them improve their teaching performances. Markow et al. (2013) found that principals agree with this notion and believe that it is their main responsibility to use data in school wide decision making. More specifically, when asked to rank the most important experience and skills for a school principal, 85% of principals surveyed indicated that a principal's use of data about student performance to improve instruction is the most important. The following section reviews literature that examines the relationship between teacher evaluation scores and student achievement outcomes.

Heneman, Milanowski, Kimball, and Odden (2006) conducted a study that evaluated the effectiveness of standards-based evaluation systems within the context of knowledge and skills-based pay for teachers. The study was done for the Consortium for Policy Research in Education (CPRE) and used four different sites, which included: the Cincinnati school system, the Vaughn Charter School in Los Angeles, the Washoe County school system in Nevada, and the Coventry school system in Rhode Island. All sites implemented a standards-based teacher evaluation system. The relationship between teachers' performance and student achievement was assessed by correlating teachers' overall evaluation scores with estimates of the value-added academic achievement of the teachers' individual students. The study also attempted to control for potential confounding variables such as prior student achievement and socioeconomic status.

Heneman et al. (2006) analyzed three years' worth of data from each site, and computed the average correlation at each site for both reading and math scores and teacher performance. Although the correlational values varied across sites, they all demonstrated positive relationships. At the Vaughn School, the authors report an average

correlation over the three years of the study of .37 in reading and .26 in math; in Cincinnati, the relationship was .35 in reading and .32 in math; in Washoe, the relationship was .22 in reading and .21 in math; and in Coventry, the relationship was .23 in reading and .11 in math.

Archibald (2006) investigated educational resources that relate positively to student achievement. The author presented four major categories of school expenditures, which included instruction, instructional support, leadership, and operations and maintenance. The current literature review focuses on the instruction category because it includes teacher evaluation and performance. The Archibald (2006) study utilized data from elementary schools in Washoe County, Nevada, that uses a standards-based teacher evaluation system. The sample included data taken from a database containing over 14,000 student records, 666 teachers, and 60 schools. The author used a three-level hierarchal linear model (i.e., students, teachers, and schools) to analyze the data. As a result, only students who could be matched to teachers, who could be matched to specific schools, were included. This decreased the overall sample to 7,601 student records, 421 teachers, and 53 schools. Student achievement scores included statewide testing data from 2002-2003 for grades 3-6. The teacher evaluation scores were taken from the district's performance-based evaluation system. Other variables included teacher education and years of experience.

The results indicated that the teacher performance scores were positively and significantly related to both student reading and math achievement at the $p < .05$ significance level (Archibald, 2006). In fact, one interesting finding is that teacher evaluation scores were the only teacher variables that related significantly to student

achievement. In contrast, teacher evaluation was a significant factor compared to both teacher education and years of experience (represented by the teacher's salary step) that did not relate to students' reading and math scores. This study's results reveals that not only are evaluation scores related to student achievement, but they may be a better criterion for teacher pay than more traditional factors of education and experience.

Kane, Taylor, Tyler, and Wooten (2010) attempted to address the issue of variability in teacher effectiveness by investigating whether teacher evaluation systems could predict which teachers positively affected student achievement scores. The authors used data collected by the Cincinnati public schools. The data included teacher evaluation scores between the years of 2000 to 2009 for a total of 2,071 teachers. The teacher evaluation scores were determined through Cincinnati's Teacher Evaluation System, which is based on Charlotte Danielson's standards-based evaluation framework (Danielson & McGreal, 2000). In addition, Kane et al. utilized student-level data in the form of standardized test scores (Ohio Achievement Test) for approximately 14,500 students in grades 3-8. The authors used the data to determine if teacher principal-rated evaluation scores differentiated between teachers who promoted high versus low student achievement. This was done by dividing teachers into quartiles based on their value-added estimates. Teachers within the upper quartile were then compared to teachers in the lowest quartile, and teachers in the upper quartile were also compared to teachers in the second quartile.

Using the value-added estimates and student achievement data allowed the authors to investigate the extent to which teacher evaluation scores predicted student achievement growth. The authors revealed that a one-point increase in teacher evaluation scores was

associated with student achievement gains of about one-sixth of a standard deviation in math and one-fifth of a standard deviation in reading. The authors translated the findings by placing them in context. For example, Kane et al. (2010) described how a student who is assigned to an upper quartile teacher would score .10 standard deviations higher in math and .13 standard deviations higher in reading when compared to a student who is assigned to a bottom quartile teacher. These findings point to the degree to which teacher evaluations by principals can identify and predict student performance, even when considering other student, school, and class-level factors.

Current Use of Principal Evaluation and Observation

The above studies presented evidence that principal evaluations can be useful within a larger evaluation system. The current dissertation study utilized principal perceptions because principals remain the core evaluator in most widely used teacher evaluation systems. The clearest example of this is Charlotte Danielson's Framework for Teaching, which was recently adopted by the New York City Department of Education (NYC DOE) in June 2013 (McAdoo, 2013).

Danielson's Framework for Teaching (2013) divides "teaching" into 22 components, which are clustered into four domains of teaching responsibility. This framework is similar to Danielson and McGreal's (2000) standards based evaluation system, which was presented earlier; however, there are some additions and changes. The four major clusters are: Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibilities (Danielson, 2013). School administrators rate each component using performance level rubrics. According to McAdoo (2013), the NYC DOE has adopted all 22 components, and it will be the responsibility of school

principals to conduct these observations. Danielson's Framework is used as part of a larger evaluation system in which teachers are ultimately rated as *Highly Effective*, *Effective*, *Developing*, or *Ineffective* on a scale from 0 to 100 points (McAdoo, 2013). In addition, the majority of points a teacher can earn is within the classroom observation component (60 points), while 20 points are based on state measures of student learning growth (i.e., improvement on state- wide standardized tests), and another 20 points are based on locally selected measures of student achievement (McAdoo, 2013).

Teachers are provided with two options for how the classroom observations will take place. Option 1 consists of one formal observation (McAdoo, 2013). A formal observation is a planned and announced observation, which also incorporates a pre-observation conference as well as a post-observation conference in order to discuss the lesson, activities, and expectations. The teacher and principal also meet post-observation to reflect and provide feedback. Following the formal observation, a minimum of three informal observations take place. The informal observations require the principal to make three, unannounced visits to the teacher's classroom for a minimum of 15 minutes per visit (McAdoo, 2013). Option 2 consists of six informal observations; however, no formal observation or pre- and post- conferences are required. Both options include a summative end of the year conference where the Danielson rubric is used to provide feedback (McAdoo, 2013).

Principals' teacher evaluation and teacher performance improvement. The articles discussed above outline the implementation, relationship, and predictive ability of sound teacher evaluation systems when placed within the context of student achievement. Taylor and Tyler (2011) approached the importance of teacher evaluation from a different

perspective. Using the economic principle of human capital, or the attributes gained by an employee through education and experience, the authors questioned whether and how teachers change when evaluated. Although the studies above provide evidence that teachers can be differentiated in terms of their performance, Taylor and Tyler (2010) attempted to find out if teacher evaluations can provide meaningful feedback to teachers to motivate them to increase their subsequent performances. The authors used only mid-career teachers from the same sample presented in the Kane et al. (2010) study reviewed above. Taylor and Tyler (2011) used this sample to estimate the extent to which a teacher's performance in the teacher evaluation process improves her performance in promoting student academic growth.

Taylor and Tyler (2010) concluded that, on average, mid-career teachers' participation in the evaluation process improved their effectiveness in promoting math achievement. In addition, the authors revealed that student performance improved during the school year in which the teacher was evaluated (Students' scores were approximately .07 standard deviations higher in math.) and in the years after evaluation (Students' scores were approximately .11 standard deviations higher in math.). This result provides indirect evidence that teachers' performance improved. The authors were not conclusive about why this change occurred only in math and not in reading achievement scores. One hypothesis that they provided is that, in contrast to learning math, students learn reading in many settings such as the home. Thus, if teachers have less influence on reading achievement variation, then one would see smaller increases in reading scores when teachers alter their practices.

An alternative explanation that Taylor and Tyler (2010) did not present is that

teachers are generally less competent in teaching math compared to teaching reading (Gallagher, 2004). As a result, when provided feedback, teachers will show greater gains in the area in which they are less competent. In addition, Gallagher (2004) found that principal evaluators were more knowledgeable about reading instruction and, as a result, gave more effective teachers higher ratings. Gallagher (2004) states that

> the lesser degree of pedagogical content knowledge for teachers and evaluators in math led to curriculum and instruction that was not highly aligned to state standards and evaluation that appears less able to distinguish among various teachers' skills in mathematics instruction (p. 103).

Gallagher (2004) explored the validity of teacher evaluation systems by investigating content area differences. He used a hierarchal linear modeling system to estimate value-added teacher effects by correlating them with teacher evaluation scores in literacy, math, and language arts as well as a composite measure of student achievement. The author used Vaughn Elementary School, which is a charter school in the Los Angeles Unified School District. The district serves approximately 1,200 students. The sample used for the study included second through fifth grade teachers ($N = 34$) and all of the students in their classrooms who had available pre- and post-test data ($N = 584$). Vaughn developed a teacher evaluation process that is a subject-specific adaptation of the work by Danielson (1996). The adaptation includes a detailed rubric for principals to use to evaluate teachers that describes a level of proficiency for up to 9 standards within each of 10 domains (lesson planning, classroom management, literacy, mathematics, language development, special education inclusion, social studies, science, art, and technology) (Gallagher, 2004).

Gallagher (2004) used these scores in subsequent data analyses as well as students' scores on the SAT-9 (Stanford Achievement Test) from both 2000 and 2001. The results showed that, overall, the teacher evaluation system had a statistically significant relationship with classroom effects. Specifically, the composite teacher evaluation scores were significantly related to value-added learning growth when comparing pre- and post-test SAT-9 sores. Furthermore, the author reported that the strength of the relationship for literacy was stronger than for math as well, which would have been anticipated from previous research. Gallagher revealed that the teacher evaluation score in literacy was the best predictor of variation in classroom effects, explaining 34% of the variation. The only other significant predictor was the composite teacher evaluation score, which explained 13% of the variation; however, math evaluation scores and language arts were not significant predictors of value-added effects. Additionally, this study confirms findings of the previously presented research by demonstrating that teacher evaluation scores are a valid predictor of student achievement and therefore can be used to identify and differentiate between effective and less effective teachers. Gallagher (2004) further identified the importance of cross content domain evaluation, because teacher evaluations in some domains (i.e., literacy) may be better predictors of student performance than teacher evaluations in other domains (i.e., math), as there appears to be some variation in how teachers are evaluated when using the same system.

Student achievement scores and how to incorporate them into a teacher evaluation system, remains a question for the new teacher evaluation system implemented by the NYC DOE. As discussed earlier, 40 of the 100 possible points earned by teachers are based on state-wide and local student assessment (McAdoo, 2013). These measures,

however, have not yet been identified as New York City has an extremely diverse student body with a wide range of abilities and learning styles. Current state-wide assessments include the Math and ELA tests; however, those tests are only given to students in grades 4-8. These grades represent only 16% of the city's teachers (McAdoo, 2013). For the elementary school teacher whose job it is to teacher a plethora of subject areas, using state and local assessments as part of her evaluation remains complicated. For example, it is unclear how a highly effective ELA teacher will be rated if his math instruction is only deemed effective based on student achievement data. In addition, Danielson's method requires multiple principal observations; however, it does not specify when these observations should take place. It remains a strong possibility that a teacher's competence, even when using a strict standards-based rubric, may vary depending on the subject area being taught during any single observation.

The articles presented above outline the importance of researching teacher evaluation systems and protocols that principals use. Results of well-developed teacher evaluation systems correlate positively with student achievement. Effective systems have proven to accurately distinguish more effective teachers from less effective teachers within the context of promoting student achievement and progress. Research in this area is vital, because evaluation systems are necessary to maintain our most effective teachers, as well as to inform our less effective teachers about ways to improve their performance to maximize every student's academic potential.

Evaluating Teacher Ineffectiveness

As discussed above, employing highly effective teachers can lead to gains in students' academic achievement (Gallagher 2002, 2004; Taylor & Tyler, 2011). This

finding has led to a public campaign for more systematic methods for evaluating teachers. Much controversy surrounds teacher evaluations, especially among teachers unions and lawmakers (Santos, 2012). The popular opinion is that new teacher evaluation systems should work to remove struggling teachers from the classroom.

Torff and Sessions (2005, 2009) presented the need for and use of teacher evaluation research in a different way. Their goal, as well as the goal of the current study, was to identify discrete causes of teacher ineffectiveness within defined contexts in order to help remediate these shortcomings. Torff and Sessions (2005, 2009) make a clear connection between principals' identification of the causes of teacher ineffectiveness and the impact that study of teacher ineffectiveness could have on teacher training and certification programs.

Another reason for focusing on causes of teacher ineffectiveness that Torff and Sessions (2005, 2009) did not discuss is the inability of previous and current evaluation systems to distinguish various degrees of teacher effectiveness (Sartain et al., 2011). According to Sartain et al. (2011), a closer examination of the current evaluation systems in the Chicago Public Schools, which primarily rely on principals' ratings of teachers, revealed that these evaluations are only able to identify causes of teacher effectiveness rather than identifying discrete causes of ineffectiveness. Sartain et al. (2011) reported that teachers were only rarely identified as Unsatisfactory (.3%) or even Satisfactory (7%). As a result, the majority (93%) of teachers were rated as Excellent or Superior according to the state's checklist evaluation system. Similarly, Markow (2013) found that a vast majority of principals surveyed said that their teachers as "excellent" (63%) and an additional 35% of principals describe their teachers as doing a "pretty good" job.

Again, it is important to note that the goal of evaluation is not to just identify a larger percentage of 'ineffective' teachers. Instead, the goal is to make the evaluation process more meaningful in order to provide information to teachers, schools, administrators, and teacher training programs that will allow for teacher growth and improvement.

Torff and Sessions (2005, 2009). The following section provides a detailed review of a study by Torff and Sessions (2005, 2009), which acted as a framework for the current study. The goal of Torff and Sessions' (2005, 2009) studies were to identify what principals believe are the most frequent causes of teacher ineffectiveness, within a high school sample, using a newly developed measure. Specifically, they investigated if there was a difference between content knowledge and pedagogical knowledge related to general teacher ineffectiveness according to secondary school principal ratings. In short, they asked the following research questions:

(1) How do principals judge the frequency with which deficiencies in content knowledge and pedagogical knowledge cause teachers' work to be ineffective?

(2) How do principals in high and low performing schools differ in judgments concerning the threats to teacher quality posed by shortages in content knowledge and pedagogical knowledge?

(3) What are subject area threats to teacher quality?

Torff and Sessions' (2005, 2009) method. First, Torff and Sessions (2005) identified 150 high performing and 150 low performing secondary schools within New York State based on data from the New York State Education Department (NYSED) website. Utilizing a need to resource capacity index, which is a measure of a district's ability to meet the needs of its students with local resources, the website classifies school

districts as having high needs, average needs, or low needs. New York State identifies high needs schools as those who meet one of the following criteria:

> (a) large urban school districts such as New York City, (b) urban-suburban districts that are above the 70th percentile in need to resource capacity (and have more than 100 students per square mile or an enrollment greater than 2, 500 and more than 50 students per square mile, (c) rural districts that are at or above the 70th percentile in need to resource capacity (and that have fewer than 50 students per square mile and an enrollment of less than 2,500) (Torff & Sessions, 2005, p. 532).

Torff and Sessions (2005) referred to the high needs schools as low performing and the low needs schools as high performing due to a reported high correlation between needs and academic performance on the NYSED website.

Utilizing these criteria for high (including low and average need schools) and low performing (including schools identified as solely high needs), Torff and Sessions (2005) randomly selected 150 schools for each criterion, totaling 300 schools for inclusion in their study.

Torff and Sessions (2009) solicited principals from 350 schools across New York State and Michigan by randomly selecting 175 schools from each state. A total of 251 principals responded. It should be noted that Torff and Sessions (2009) did not divide their sample into low and high performing schools. In both Torff and Sessions' (2005, 2009) studies, the authors sent principals of each school a survey, which included postage paid return envelopes and directions.

Torff and Sessions' (2005, 2009) survey instrument. Torff and Sessions (2005,

2009) developed their survey instrument by examining 20 teachers' guides created by administrators at 20 different school districts across New York State. These guides were investigated from both high and low needs districts (10 from each) as defined above. Investigation of these guides revealed similarities within their descriptions of the knowledge and skills that are expected of teachers. According to Torff and Sessions (2005, 2009) all of the guides included criteria that fit into the following five dimensions:

(1) Content Knowledge – suitable expertise in the subject being taught

(2) Lesson-Planning Skills – preparation of appropriate learning experiences prior to an instructional period

(3) Lesson-Implementation Skills – effective execution of planned learning experiences during an instructional period

(4) Ability to Establish Rapport With Students – adequate human relations and communication skills

(5) Classroom Management Skills – ability to successfully keep students on task and attentive (p. 532)

Torff and Sessions (2005) then used these five selected dimensions to develop their survey instrument. The dimensions were treated as dependent variables. Principals rate the frequency of each dimension in relation to how frequently they judge that dimension to be associated with teacher ineffectiveness on a 4-point scale (*very rarely* to *frequently*). Five independent variables were also included, which acted as covariates. These included: principal age, gender, years of experience as a classroom teacher, years of experience as an administrator, and educational attainment.

The Torff and Sessions (2009) study included ratings across five subject areas:

English, Math, Social Studies, Science and Foreign Language.

Torff and Sessions' (2005) results. In total, 242 principals (169 men and 73 women) of secondary schools (112 in low performing schools and 130 in high performing schools) completed the survey. Most (198) had attained a Master's Degree plus 30 credits, followed by 32 who had a doctoral degree, and 12 principals with a Master's Degree (Torff & Sessions, 2005).

Torff and Sessions (2005) used Hotelling's T^2 analysis to simultaneously compare the means of the five dependent variables. The results revealed that the average frequency of content knowledge as a cause of teacher ineffectiveness was significantly lower ($p < .0001$) than the average frequencies for the other four dimensions (Torff & Sessions, 2005). In sum, Torff and Sessions (2005) reported that the most frequent causes of teacher ineffectiveness could be placed in three categories, classroom management skills, lesson-implementation and skills, and rapport with students. These were followed by lesson planning skills and content knowledge, the least frequent cause of teacher ineffectiveness. Torff and Sessions (2005) also found that principals' perceptions regarding the causes of teacher ineffectiveness were mainly stable across types of schools and participant demographics.

Torff and Sessions' (2009) results. In total, 251 high school principals across New York State and Michigan completed the survey. They conducted the same types of statistical analyses as used in Torff and Sessions (2005). The authors report that when data were aggregated from the five subject area ratings, classroom management was the greatest perceived threat to teacher quality and content knowledge was the least likely cause of perceived teacher ineffectiveness. In addition, the authors report little variation

across subject area ratings. As a result, Torff and Sessions (2009) concluded that principals judged the causes of teacher ineffectiveness to be similar across the five secondary subject areas.

Pilot Study

The following sections provide a review of the pilot study that I conducted based on the general framework of Torff and Sessions' (2005) study as well as information gathered in the literature review. A summary of the methods, results, and extensions that were incorporated into the present dissertation are included. In addition, data from the pilot also aided in developing the dissertation hypotheses.

Participants. Participants for the pilot study included 14 elementary school principals whose names and addresses appeared on the New York State Education Department (NYSED) website, which provides data for all New York State public schools. I used data from the NYSED website for the both the pilot study and the current study. The information included names of public schools as well as statewide student achievement data in English Language Arts (ELA) and Math, specifically mean scores for grades 3 and 5 from both the state Math and ELA tests. Names of principals for the pilot were taken directly from this list in the same order as they were presented on the website. The researcher did not randomly choose names in order to maximize the potential number of participants.

Measure. The measure utilized in the pilot, and the dissertation is the survey instrument that Torff and Sessions (2005) used. The authors gave me permission to reproduce and distribute the survey (See Appendix D). The instrument identifies five dimensions of teacher ineffectiveness discussed above. On the survey, the dimensions

are expressed follows (Torff & Sessions, 2005, p. 532):

1. Fails to demonstrate needed content knowledge
2. Fails to write effective lesson plans
3. Fails to implement lesson plans skillfully
4. Fails to establish sufficient rapport with students
5. Fails to maintain satisfactory classroom discipline

Participants rate the frequency of each dimension in relation to how frequently they judge that dimension to be associated with teacher ineffectiveness on a 4-point scale (*very rarely* to *frequently*). Principals perform this rating three times. The first is a rating of general teacher ineffectiveness. Following this, principals rate the dimensions association with teacher ineffectiveness within the context of specific content areas (i.e., ELA and Math). In addition, the measure includes five independent variables that are used as possible covariates. These include: principal age, gender, years of experience as a classroom teacher, years of experience as an administrator, and educational attainment (Torff & Sessions, 2005).

Statistics. First, I generated descriptive statistics for the variables. Descriptive statistics included means, standard deviations and confidence intervals. This was first done for the general teacher ineffectiveness ratings, and then for the ELA and Math content area specific teacher ineffectiveness ratings.

Results. Table 1 reports descriptive statistics for the 14 principals.

Table 1

Descriptive Statistics for Participants' Age, and Years Teaching and in Administration

Demographic	*M*	*SD*
Age	48.36	9.00
Years Teaching	10.43	7.65
Years as an Administrator	10.57	6.16

Note. $N = 14$.

The above statistics describe the sample of participants used in the pilot. The sample appeared to be an experienced group in terms of both teaching and working as an administrator. Furthermore, no principal had less than a Master's degree plus 30 credits, and the data set consisted of 11 females and 3 males.

Table 2 presents the principals' ratings of the perceived causes of general teacher ineffectiveness. The table reports the order of elementary school principals' ratings from the least frequent to the most frequent perceived causes of teacher ineffectiveness. Table 2 shows that, with the exception of Rapport with Students, that Content Knowledge was rated lower than the other dimensions. Torff and Sessions (2005) also found that high school principals rated Content Knowledge lower that the other dimensions of teacher ineffectiveness.

Table 2

Differences between Content Knowledge and the Four Pedagogical Dimensional Causes of General Teacher Ineffectiveness (Rapport, Lesson Planning, Classroom Management, and Lesson Plan Implementation)

Dimension	*M*	*SD*	*Confidence Interval 95%*
Content Knowledge	2.64	0.75	2.25-3.03
Rapport with Students	3.00	0.96	2.50-3.50
Lesson Planning Skills	3.29	0.73	2.91-3.67
Classroom Management	3.43	0.65	3.09-3.77
Lesson Plan Implementation Skills	3.50	0.65	3.16-3.84

Note. $N = 14$.

Table 3 presents elementary school principals' ratings of the causes of teacher ineffectiveness for ELA in order from least to most frequent. Principals rated Content Knowledge significantly lower than Lesson Plan Implementation Skills as a perceived cause of ELA teacher ineffectiveness. There were no other differences in principals' ratings of the dimensions. Tables 2 and 3 also indicate that the ordering of principals' perceptions of general teacher ineffectiveness is closely related to principals' ratings of teacher ineffectiveness for reading and other ELA tasks.

Table 3

Differences between Content Knowledge and the Four Pedagogical Causes (Rapport, Lesson Planning, Classroom Management, and Lesson Plan Implementation) of Teacher Ineffectiveness within ELA

Dimension	*M*	*SD*	Confidence Interval 95%
Content Knowledge	2.64	0.93	2.15-3.13
Rapport with Students	2.79	0.98	2.19-3.21
Lesson Planning Skills	3.07	0.83	2.64-3.50
Classroom Management	3.29	0.83	2.86-3.72
Lesson Plan Implementation Skills	3.36	0.75	2.97-3.75

Note. $N = 14$.

The following analysis reveals how principal perceptions are different than those for general and ELA teacher ineffectiveness when principals are asked to identify perceived causes of teacher ineffectiveness in regard to Math. As mentioned, principal ratings are consistent across general causes as well as causes of teacher ineffectiveness within the context of ELA. In both instances, principals rated Content Knowledge as a significantly less frequent cause of teacher ineffectiveness than the other dimensions, in particular Lesson Plan Implementation. The following table demonstrates how this trend differs when principals rated the causes of teacher ineffectiveness in Math. Table 4 reports the order of ratings from the least frequent to the most frequent principal perceived causes of teacher ineffectiveness, within Math endorsed by elementary school principals in the pilot sample.

Table 4

Differences between Content Knowledge and the Four Pedagogical Causes (Rapport, Lesson Planning, Classroom Management, and Lesson Plan Implementation) of Teacher Ineffectiveness within Math

Dimension	*M*	*SD*	*Confidence Interval 95%*
Content Knowledge	3.29	0.83	2.86-3.72
Rapport with Students	2.79	0.89	2.32-3.26
Lesson Planning Skills	3.29	0.73	2.91-3.67
Classroom Management/ Discipline	3.36	0.83	2.93-3.79
Lesson Plan Implementation Skills	3.50	0.65	3.16-3.84

Note. $N = 14$.

This table demonstrates how principals' perceptions of the cause of teacher ineffectiveness vary when framed within the context of Math. Content Knowledge is no longer the least most frequent cause of teacher ineffectiveness. When comparing Tables 3 and 4, readers should note that principals differed in their view of the causes of teacher ineffectiveness in math relative to their view of the causes of teacher ineffectiveness in ELA.

Discussion. In general, the pilot results and those of Torff and Sessions (2005, 2009) showed that principals rated Content Knowledge as a less important cause of teacher ineffectiveness than the other four factors. This result implies that principals believe that teacher ineffectiveness is not caused by a lack of content knowledge, but is instead caused by a lack of pedagogical knowledge. The pilot results suggest, however,

that whether principals perceive content knowledge as a cause of teacher ineffectiveness or not may depend on the subject matter for elementary school principals.

This information could inform future research and policy. For example, if differences do exist in principals' ratings of teacher ineffectiveness that are dependent on academic content, then the evaluation of teachers should emphasize certain criteria dependent on academic subject area. This would mean that principals' evaluation of teachers should be content specific and not global in nature. Furthermore, this research could also inform elementary teacher preparation programs. If principals identify causes of teacher ineffectiveness in dimensions not emphasized in elementary teacher preparation programs, then these programs may supplement or amend their curricula. Curriculum amendment may also depend on content area. For example, results of this study may indicate that more content knowledge instruction is needed in math, however, lesson-implementation should be emphasized when preparing teachers to teach ELA.

Purpose

Readers should note that, although the present study only investigated principals' ratings of elementary school teachers' ineffectiveness, this is not to suggest that principals' ratings should be the only criterion on which teachers are assessed. The dissertation author supports multi-faceted, comprehensive teacher evaluations in keeping with the school psychology tradition of evaluation (Wilkerson et al., 2000). Principals' ratings of teachers are one just part of a comprehensive teacher evaluation system.

The current study is a replication and extension of Torff and Sessions' (2005, 2009) studies that measured principals' perceptions of the causes of teacher ineffectiveness within high school classrooms. More specifically, their studies investigated whether

principals believed that deficiencies in content knowledge or pedagogical knowledge cause teachers' ineffectiveness. The present study extended the research by Torff and Sessions (2005, 2009) to elementary school principals as well as investigated if principal-rated teacher ineffectiveness differences existed between specific, academic content areas (ELA and Math). Pilot results indicated that principals may view the reasons for teacher ineffectiveness differently across these two content areas.

Hypotheses

The current study proposed the following hypotheses:

HO1: Consistent with Torff and Sessions' (2005) findings, general causes of elementary school principals' perceived teacher ineffectiveness will be (in order from the most frequent causes to least frequent causes): classroom management skills (pedagogical), lesson- implementation skills (pedagogical), rapport with students (pedagogical), lesson-planning skills (pedagogical), and content knowledge.

HO2: Consistent with Torff and Sessions (2009) elementary school principals' perceived causes of teacher ineffectiveness within specific content areas (Math and ELA) will not differ from each other.

H03: Consistent with Torff and Sessions' (2009) findings, elementary school principals' perceived causes of teacher ineffectiveness, within the context of ELA and Math instruction, will be (in order from the most frequent causes to least frequent causes): classroom management skills (pedagogical), rapport with students (pedagogical), lesson- implementation skills (pedagogical), lesson-planning skills (pedagogical), and content knowledge.

CHAPTER III

Method

The following section presents the methods used in the current study. More specifically, the section describes the recruitment and description of participants, the survey instrument, the data collection process, and a description of the statistical analyses.

Participants

Recruitment. Participants include elementary school principals within New York State. Similar to procedures used by Torff and Sessions (2005), the researcher obtained school information by accessing the New York State Education Department (NYSED) website. This website includes school level data for all New York State public schools. The researcher used this site to obtain the names of public elementary schools.

Once he had identified a school, the researcher conducted an Internet search to obtain the name and email address of the school's principal. He did this procedure for each elementary school in the state. An email (see Appendix B) was sent to each principal, including a link to the consent form (see Appendix C) and survey instrument (see Appendix A) located on Survey Monkey. On the site, the principals were able to consent to participation as well as complete the survey on-line.

A total of 1,255 surveys were emailed to individual elementary school principals in New York State. Out of that pool of 1,255 possible participants, a total of 97 principals responded, resulting in a response rate of 12.81%. In addition, web links to the survey were also disseminated via word of mouth. This process entailed sending links to colleagues as well as university professors who then passed the link to elementary school

principals whom they knew. This process resulted in an additional 7 surveys being completed, bring the total number of individuals who responded to the study solicitation to 104.

Participants' obtained and missing data. Of the 104 respondents, 83 (79.81%) completed the entire survey (i.e., consent, demographic information, and all survey questions that concerned both general and content area specific causes of teacher ineffectiveness). An additional 7 (6.73%) respondents provided consent and demographic information, and completed the first part of the survey concerning their beliefs about the general causes of teacher ineffectiveness. Thus, a total of 90 respondents (86.54% of those who responded) were included in the sample for data analysis and hypothesis testing for the general ratings of teacher ineffectiveness.

An additional 10 (9.62%) respondents only completed consent and demographic information but did not complete any survey questions. Finally, 3 (2.88%) respondents provided consent, but did not complete demographic or survey questions, and thus, provided no data for the study. An additional respondent's data were excluded because he worked as a principal in Hawaii.

Description of participants. The following tables present demographic information for respondents who participated in at least the general ratings of teacher ineffectiveness portion of the survey (see Table 5 and 7) and those who provided demographic information but did not complete the survey (see Table 6 and 8). Tables 5 and 6 represent continuous variables (Age, Years of Teaching, and Years as an Administrator). Tables 7 and 8 present categorical variables (Educational Attainment and Gender) in the form of one-way frequency tables.

Table 5

Descriptive Statistics for Principals Who Completed the Survey (Continuous Variables)

Demographic	*M*	*Min*	*Max*	*SD*
Age[a]	50.59	36	72	7.89
Years Teaching[b]	11.69	1	30	6.02
Years as an Administrator[b]	12.37	4	34	6.12

[a]$N = 90$. [b]$n = 89$.

Table 6

Descriptive Statistics for Principals Who Did Not Complete the Survey (Continuous Variables)

Demographic	*M*	*Min*	*Max*	*SD*
Age	45.80	35	56	8.39
Years Teaching	10.90	5	20	4.41
Years as an Administrator	9.85	1	22	7.03

Note. $N = 10$.

Table 7

One-Way Frequency Table for Principals Who Completed the Survey (Categorical Variables)

Variable	*Frequency*	%
Gender		
Male	27	30
Female	63	70
Educational Attainment		
Bachelor's	0	0
Master's	6	6.67
Master's + 30	74	82.22
Ph.D.	10	11.11

Note. $N = 90$.

Table 8

One-Way Frequency Table for Principals Who Did Not Complete the Survey (Categorical Variables)

Variable	*Frequency*	%
Gender		
Male	5	50
Female	5	50
Educational Attainment		
Bachelor's	0	0
Master's	0	0
Master's + 30	10	100
Ph.D.	0	0

Note. $N = 10$.

An eye-ball comparison of the two groups indicates that, on average, the group that completed the survey was more experienced and older. The average age for principals who completed the survey is 50.59 years old versus 45.80 years for those who did not complete the survey. In addition, the survey completion group has approximately one year more experience teaching (11.69 vs. 10.90) and approximately three more years of experience as an administrator (12.37 vs. 9.85). Tables 8 and 9 show that the majority of principals had their Master's + 30 credits (82.2% of completers vs. 100% of non-completers). It should be noted, however, that these differences could not be tested

statistically because the sample size for the non-survey completion group was quite small ($n = 10$).

Instrument

The survey instrument used in the current study (see Appendix A) is the same instrument used in the study by Torff and Sessions (2005, 2009). Principals rate reasons for their perceived causes of teacher ineffectiveness according to the following five dimensions:

(1) Content Knowledge – suitable expertise in the subject being taught

(2) Lesson-Planning Skills – preparation of appropriate learning experiences prior to an instructional period

(3) Lesson-Implementation Skills – effective execution of planned learning experiences during an instructional period

(4) Ability to Establish Rapport With Students – adequate human relations and communication skills

(5) Classroom Management Skills – ability to successfully keep students on task and attentive (Torff & Sessions, 2005, p. 532)

Principals rate each dimension in relation to how frequently they judge that dimension to be associated with perceived teacher ineffectiveness on a 4-point scale (*very rarely* to *frequently*) where higher ratings demonstrate greater perceived threats to teacher quality. More specifically, the survey asked the respondents to rate the frequency with which each dimension represents a significant contribution to teacher ineffectiveness. Principals rate these dimensions three times. The first is a rating for general teacher ineffectiveness. Following this, principals rate the Dimensions association with teacher

ineffectiveness within the context of specific content areas (i.e., ELA and Math). The survey also asks for principal age, gender, years of experience as a classroom teacher, years of experience as an administrator, and educational attainment.

In addition, the questionnaire asks for the following school wide factor data: grade levels in the school, location and region of the school, average class size, percentage of students who receive free or reduced lunch and the percentage of students with an IEP. Finally, the questionnaire includes qualitative questions to obtain supplemental information from principals. These questions provided principals with a place to express their opinions regarding teacher evaluation systems at their school. The items ask for information that comprises the evaluations, principals' opinion of their evaluation system, information that they would like to include in teacher evaluations, and ideally, how they would wish to do evaluations (see Appendix A).

Using this questionnaire, Torff and Sessions (2005, 2009) found that, contrary to some researchers' beliefs (Gross, 1999; Ravitch, 2000), principals did not view teacher ineffectiveness as due to lack of content knowledge. Instead, New York State high school principals cited pedagogical deficiencies (i.e., classroom management, lesson implementation, student rapport) as threats to teacher quality.

Procedure

Torff and Sessions (2005, 2009) granted the researcher permission to use the survey instrument and provided him with an electronic version of the survey instrument to be used in the current study. In addition, Torff and Sessions provided the researcher a supplemental survey that asked the same questions; however, the supplemental survey framed the questions within the context of specific academic content areas (i.e., ELA and

Math). The researcher then transcribed the survey instrument, which included both the general and specific academic content items, into an electronic version using Survey Monkey.

Following the transcription of the survey, the researcher conducted Internet searches to obtain a list of New York State elementary school principals. The searches began by finding schools listed on the NYSED website. Once the researcher identified an elementary school, he conducted an Internet search to find the school's principal's name and email address. In this manner, the researcher compiled a list of principals' emails and saved it to Survey Monkey. A link to the questionnaire was then emailed to principals via Survey Monkey along with a brief description of the study (see Appendix B).

In addition to principals' ratings of perceived causes of teacher ineffectiveness, the researcher also obtained student achievement data through the NYSED website. These data are in the form of mean scaled scores for both ELA and Math for 3rd through 8th grade students for each of the participating principals' schools. In addition, the website provides information on percentages of students per grade who score within one of the four possible levels (Level 1 - 4). The following are descriptions of each of the four levels as presented on the NYSED website ("Descriptions of Performance Levels", 2013):

Level 1: Below Standard

Student performance does not demonstrate an understanding of the English language arts/Mathematics knowledge and skills expected at this grade level.

Level 2: Meets Basic Standard

Student performance demonstrates a partial understanding of the English language arts/Mathematics knowledge and skills expected at this grade level.

Level 3: Meets Proficiency Standard

Student performance demonstrates an understanding of the English language arts/ Mathematics knowledge and skills expected at this grade level.

Level 4: Exceeds Proficiency Standard

Student performance demonstrates a thorough understanding of the English language arts/Mathematics knowledge and skills expected at this grade level.

Not every school included in the survey had grades 3-8. As a result, subsequent analyses included only the data for those grades for each specific school.

Statistical Analyses

Descriptive statistics provided totals, means, minimum values, maximums values, standard deviations, effect sizes, and ranges for the various dependent and independent variables (i.e., percentage of students who have a IEP, percentage of students who receive free and reduced lunch and average class size) as well as the five Dimensions of potential teacher ineffectiveness (i.e., Lesson-Implementation Skills, Ability to Establish Rapport with Students, Classroom-Management Skills, Lesson-Planning Skills, and Content Knowledge). In addition, the researcher conducted correlational analyses to determine strength and direction of study variables.

In Torff and Sessions' (2005, 2009) studies, following descriptive analyses, a Hotelling's T^2 analysis was done to simultaneously compare the means of the dependent variables (Content Knowledge, Lesson Planning Skills, Lesson Implementation Skills, Rapport With Students and Classroom Management). The current dissertation instead

utilized an ordinal probit model to analyze dependent variables. This model allowed the researcher to conduct likelihood ratio testing to determine if the full model (Dimension X Domain) is preferred to the reduced model (Dimension only). The analysis addressed whether or not principal ratings varied depending instructional content area (Domain). In addition, general pairwise comparisons were conducted to determine statistical differences between the Dimension ratings.

For the present study, the Full Model includes the variable of Dimension (i.e., the potential perceived causes of teacher ineffectiveness: Content Knowledge, Implementation of Lesson Plans, Writing Lesson Plans, Rapport with Students, and Classroom Management) as well as Domain (ELA and Math ratings). The Reduced Model only includes the variable of Dimension. Table 9 presents the Full and Recued Models.

Table 9

Ordinal Probit Model

Reduced Model (Dimension)	Full Model (DimensionXDomain)	
Dimension	Dimension	Domain
Content Knowledge	Content Knowledge	ELA
Writing Lesson Plans	Writing Lesson Plans	Math
Implementing Lesson Plans	Implementing Lesson Plans	
Classroom Management	Classroom Management	
Rapport With Students	Rapport with Students	

The researcher conducted likelihood ratio tests to determine if statistical differences existed between models. Finally, a mixed ordinal probit model followed by pairwise comparisons was used to assess the principals' ratings of the five Dimensions across Domains (i.e., General, ELA, and Math). The comparisons allowed the researcher to assess principal perceptions across Domains because the mixed ordinal probit model adjusts for individual principals' propensity to use the scale in different ways.

Qualitative Procedure and Analysis

The final four questions on the survey (Questions 14-17) are open-ended (See Appendix A). The questions allowed for principals to expand on their responses and express their opinions. The researcher reviewed every survey and identified common themes in participants' responses. Following the identification of themes, the researcher created categories to use to code each of the participants' responses (See Appendix I). It should be noted that each individual question to which a participant did not respond was coded as 0. Coding was then independently completed by both the researcher and a research assistant. The research assistant is a former teacher, who received his master's plus 60 credits in education. The research assistant also has experience working in school administrative roles, which was deemed useful for the current study.

The researcher provided the assistant training on how to code each response that first included a review of each of the category definitions for each question. During category review, the assistant asked clarifying questions that particularly related to terminology used in the definitions. Once all questions were answered and the definitions were reviewed, the researcher and research assistant independently coded five participants' responses, compared results, and discussed any coding inconsistencies.

Once they established perfect coding agreement on the responses of these five principals, they moved onto coding the remaining principals' responses. Following the coding of each response, the research assistant and research reviewed their coding and discussed all inconsistencies and mutually decided upon the most appropriate coding category to use. Following the coding of all responses, the researcher then calculated frequencies and percentages for responses to each question (see Tables 21-24 in Chapter IV) as well as the percentage of each response category per question for each coder (see Appendix J) separately. Inter-rater reliability was calculated by computing Cohen's Kappa.

Chapter IV

Results

The following section presents the results. The chapter begins with descriptive statistics and correlational analyses. Following this, each hypothesis is tested. The next section presents coded participants' answers to the four open-ended items regarding teacher evaluations at their respective schools.

Descriptive Statistics and Correlations of Study Measures

Table 10 and 11 below presents descriptive statistics for school and classroom level variables for the current study. These variables include average class size, percentage of students who have an IEP, percentage of students who receive free or reduced lunch, location of school, grade levels, and mean scaled scores for New York State ELA and Math tests. (Tables 5 through 7 in the Method section contain principals' demographic information.)

Table 10 is a one-way frequency table that summarizes the categorical variables: location of school and grade level. The majority of schools represented (63.3%) is Pre-k or K-5 and the second most represented grade level is Pre-k or K-6 (11.11%). In addition, the majority of schools come from either Urban (46.67%) or Suburban (37.78%) locations with Rural areas making up 15.56% of the sample.

Table 10

One-Way Frequency Table for Location of School and Grade Levels (Categorical Variables)

Variable	*Frequency*	%
Grade Level		
Pre-k(K)- 2nd Grade	2	2.22
Pre-k(K)- 4th Grade	5	5.56
Pre-k(K)- 5th Grade	57	63.33
Pre-k(K)- 6th Grade	10	11.11
Pre-k(K)- 8th Grade	8	8.89
Pre-k(K)- 12th Grade	5	5.56
6th- 8th Grade	2	2.22
6th- 8th Grade	1	1.1
Location of School		
Rural	14	15.56
Suburban	34	37.78
Urban	42	46.67

Note. $N = 90$.

Table 12 presents the descriptive statistics for classroom level variables, which include average class size, percentage of students who receive free or reduced lunch, percentage of students who have an IEP, and mean scaled scores for NYS ELA and Math test results. In the current study, the percentage of students who receive free or reduced

lunch is 50%. New York State reports a percentage of 48.10% (State Education Data Profiles, 2013) for all schools. In addition, the National Center for Education Statistics reports that the statewide percentage of students who have an IEP is 16.55%, which is similar to the 15% of the current sample with IEPs (State Education Data Profiles, 2013). Finally, the current sample reports an average class size of 23.35 students. A direct comparison to the state-wide class size average is not possible because the state reports a pupil/teacher ratio (12.92 students per teacher) (State Education Data Profiles, 2013) that is not comparable to average class size, because many classrooms may have multiple teachers and/or teacher assistants.

Table 11

Descriptive Statistics for School Level Variables (Continuous Variables)

Variable	*M*	*Min*	*Max*	*SD*
% Students with an IEP	15	0	100	14.25
% Students who receive F/R Lunch	50	1	100	34.68
Average Class Size	23.35	8	50	4.92

Note. $N = 90$.

The final descriptive statistic is student achievement as measured by NYS Math and ELA mean scaled scores. Of the total sample of 90 schools, 86 reported school wide test scores. The four that did not report scores consisted of three schools that only house

students up to second grade who do not participate in statewide testing (The NYS ELA and Math standardized tests begin in Grade 3.), and one school that is entirely for special education with students do not participate in the NYS standardized assessment. Table 12 below presents the descriptive statistics for mean scaled scores for Math and ELA.

Table 12

Descriptive Statistics for Student Achievement Data (Continuous Variables)

Variable	*M*	*Min*	*Max*	*SD*
ELA	667.89	527.67	696.67	18.01
Math	686.34	527.33	713.67	21.19

Note. $n = 86$.

The sample of schools in the current study appears to be representative of New York State schools in regard to student achievement on the state wide standardized testing. NYS reports mean scaled scores of 665.67 for ELA (compared to 667.89 in the current study) and 684.33 for Math (compared to 686.34 in the current study) (English Language Arts and Mathematics Assessment Results, 2013).

In summary, the descriptive statistics indicate that the current study's sample is similar to elementary schools in New York State in regard to average percentage of students with an IEP, percentage of students who receive free or reduced lunch, and student achievement in regard to results on state-wide standardized testing. These data allow for some limited generalization of current results to the larger estimated population of New York State elementary school principals.

Table 13 presents correlations using school and principal level demographic variables. Results suggest that these variables are highly related. This is particularly true for Principals' Age, because it demonstrates strong relationships with Years as an Administrator and Years Teaching.

Table 13

Correlations among Principal Demographic Variables

	Age	Yrs. Teaching	Yrs. as an Administrator	Education Level
Age	1			
Yrs. Teaching	.44**	1		
Yrs. Admin.	.59**	.08	1	
Ed. Level	-.20	-.12	-.04	1

*Note. N=90. **p < 0.01*

Hypothesis Testing

Hypothesis 1. Hypothesis 1 states that, consistent with Torff and Sessions' (2005) findings, General causes of teacher ineffectiveness will be (in order from the most frequent causes to least frequent causes): Classroom Management/Discipline Skills (pedagogical), Lesson-Implementation Skills (pedagogical), Rapport with Students (pedagogical), Lesson-Planning/Writing Lesson Plan Skills (pedagogical), and Content Knowledge.

To address this hypothesis, the researcher first calculated descriptive statistics for the five Dimensions across the General causes of teacher ineffectiveness Domain (See Table 14). In addition to means and standard errors, Hedges *g* was also calculated in

order to determine the effect sizes for the difference in ratings between Math and ELA ratings. Calculated means indicate a different ordering of perceived causes of teacher ineffectiveness by elementary school principals than those reported by high school principals in Torff and Sessions (2005).

Table 14

Means, Standard Errors, and Effect Sizes for Survey Responses

Dimension	Content Knowledge *M* (*SE*)	Writing Lesson Plans *M* (*SE*)	Implementing Lesson Plans *M* (SE)	Rapport With Students *M* (*SE*)	Classroom Mana. *M* (*SE*)
Domain					
General[a]	2.80 (.092)	3.18 (.083)	3.37 (.076)	3.02 (.087)	2.97 (.083)
ELA[b]	2.82 (.099)	3.14 (.097)	3.24 (.076)	2.90 (.080)	2.95 (.095)
Math[b]	3.04 (.090)	3.13 (.095)	3.27 (.073)	2.92 (.081)	2.94 (.096)
Hedges *g*	0.24	-0.01	0.04	0.02	-0.01

Note. Principal ratings of teacher ineffectiveness were on a four-point scale with higher numbers indicating greater association of the dimension with teacher ineffectiveness.
[a]$N = 90$, [b]$n = 83$.

Table 16 shows that Content Knowledge is the only cause of teacher ineffectiveness that both elementary and secondary school principals ranked in the same order (i.e., last). Elementary school principals in the current study ranked Implementing Lesson Plans as the greatest cause of teacher ineffectiveness, but high school principals in Torff and Sessions' study ranked Discipline as the greatest cause of teacher ineffectiveness.

Table 15

Comparison of Perceived Causes of Teacher Ineffectiveness in the Current Study and in the Study by Torff and Sessions (2005) - Most Frequent to Less Frequent

Torff and Sessions (2005)	Current Study
Discipline Skills	Implementing Lesson Plans
Implementing Lesson Plans	Writing Lesson Plans
Rapport with Students	Discipline Skills
Writing Lesson Plans	Rapport with Students
Content Knowledge	Content Knowledge

Thus, HO1 is not supported, because the ordering of perceived elementary school principals' ratings of General causes of teacher ineffectiveness are different than the ordering of ratings obtained from Torff and Session's (2005) high school principal sample.

Tests of differences in principal ratings in the current sample. To determine if there are statistical differences within the current rankings of the five Dimensions ANOVA analyses were conducted. Results indicate a statistically significant difference among the Dimensions within the General ineffectiveness ratings, F (4, 444) = 6.63, $p <$ 0.001, as well as within the ELA, F (4, 410) = 3.79, $p < 0.01$, and Math, F (4, 410) = 2.72, $p < 0.05$, ineffectiveness ratings. This indicates that there are differences between the frequency of perceived causes of teacher ineffectiveness among principals. Appendix E presents the complete ANOVA tables for the analyses.

To assess further differences between ineffectiveness Dimensions, the researcher

conducted a series of pairwise comparisons (See Appendix F) to determine possible significant mean differences in principals' perceived causes of teacher ineffectiveness. The results confirm that Implementing Lesson Plans was by far the most frequently perceived cause of teacher ineffectiveness, because principals rated it significantly higher than all other Dimensions with the exception of Writing Lesson Plans. Specifically, Implementing Lesson Plans was different from Discipline and Rapport with Students at the $p < 0$ .05 level, and from Content Knowledge at the $p < .001$ level. Another significant difference was found between Writing Lesson Plans and Content Knowledge ($p < .05$).

Hypothesis 2. Hypothesis 2 states that consistent with Torff and Sessions (2009) elementary school principals' perceived causes of teacher ineffectiveness within specific content areas (Math and ELA) will not differ from each other.

The first step in testing this hypothesis was to calculate descriptive statistics for each dimension across the specific content Domains of ELA and Math (See Table 15 above). Compared to General causes of teacher ineffectiveness, principals' mean rating scores of ineffectiveness within ELA are the same for three of the five dimensions. Thus, Implementing Lesson Plans and Writing Lesson Plans were endorsed as the first and second most frequent perceived cause of teacher ineffectiveness for both General and ELA, and Content Knowledge was the least frequently perceived cause. Shifting of rating order was only observed between Classroom Discipline (ranked third in General and ranked fourth in ELA) and Rapport with Students (ranked fourth in General and ranked third in ELA).

Within the Math domain, similar to the General and ELA domains, Implementing

Lesson Plans and Writing Lesson Plans were the dimensions that had the highest mean rating scores. A notable difference for Math, compared to General and ELA ratings, was that Content Knowledge was not the least frequently endorsed cause of teacher ineffectiveness. In fact, it was ranked as the third most frequent cause of Math teacher ineffectiveness, and Rapport with Students was ranked fifth. Overall, however, mean scores indicate little difference among ratings across content area Domains with the possible exception of Content Knowledge in the Math Domain. Further statistical testing was done to determine if this was a significant change.

In addition, the adjusted R^2 values were calculated to determine how much of the variance in principals' ratings of teacher ineffectiveness can be accounted for by ELA and Math. Results show that for both ELA and Math, the amount of variance accounted for are strikingly small (ELA: adjusted R^2 = .03 and Math: adjusted R^2 = .02). These results provide evidence that, despite differences in ratings of teacher ineffectiveness within each content area Domain, the effect of the Domain (either ELA or Math) on ratings is negligible. Finally, an ANOVA analysis was run in order to compare the two content Domain ratings (See Appendix G). Once again, there was no significant difference between principals' perceptions when comparing Math and ELA ratings, F (1, 738) = .03, p > .05.

To further test this hypothesis, an ordinal probit model was used in order to conduct a likelihood ratio test between the Full (Dimension X Domain) and Reduced Models (Dimension). A chi-Square analysis was calculated to determine if the Full Model fit the data significantly better than the Reduced Model. Results indicate that the Full Model does not provide any statistically significant improvement relative to the Reduced Model,

χ^2 (1,246) = 9.54, $p > .05$. Thus, the variable of Domain (ELA and Math) does not explain any statistically significant variation in principal perception ratings.

Overall, Hypothesis 2 was not supported. Although there were some differences in principals' ratings of the Dimensions of teacher ineffectiveness within each content Domain, those ratings were similar to those found for the General ratings of teacher ineffectiveness. In addition, ANOVA and likelihood ratio testing found that ELA and Math content areas (i.e., Domains) provided little predictive value and can be interpreted as having little to no effect on principals' ratings of teacher ineffectiveness for the current sample.

Hypothesis 3. Consistent with Torff and Sessions' (2009) findings, elementary school principals' perceived causes of teacher ineffectiveness, within the context of ELA and Math instruction, will be (in order from the most frequent causes to least frequent causes): classroom management skills (pedagogical), rapport with students (pedagogical), lesson- implementation skills (pedagogical), lesson-planning skills (pedagogical), and content knowledge.

To test this hypothesis, mean and standard errors were calculated from principal ratings across the Domains (see Table 14). As described earlier, there are some principal ranking differences between Domains, because principals perceived Content Knowledge as the third most frequent cause of perceived teacher ineffectiveness in Math, compared to the least frequent cause in ELA. This difference, however, was not statistically significant as demonstrated by results of ANOVA and likelihood ratio testing.

To determine a final rank order of perceived teacher ineffectiveness among principals, a mixed ordinal probit model followed by pairwise comparisons was used to

assess the principals' ratings of the Dimensions across Domains (See Appendix H for full output). The mixed model adjusts for individual principals' propensity to use the scale in different ways. The following equation represents the mixed ordinal logistic regression where Y_{ijk} is the latent response for the i^{th} person on the jk^{th} domain:

$$Y_{ijk} = \beta_j\, dimension_j + \delta_k\, domain_k + (\beta\delta)_{jk}\, dimension_j\, X\, domain_k + \mu_I + \varepsilon_{ijk.}$$

The result of the analysis was substantial (between principal estimated variance = 0.55, 95% CI = [0.34, 0.87]), indicating considerable variability in principals' scale usage. Appendix K contains the full model output.

To determine which Dimensions principals preferred, pairwise differences were computed from the model among the Dimensions. The ordinal probit pairwise differences are naturally scaled on a standard score metric by the way the model is constructed. Table 18 contains these difference and their associated Sidak-corrected 95% confidence intervals based on Huber-White robust standard errors. The Sidak correction was used because Type 1 errors are more likely to occur as more tests are performed on a set of data (Abdi, 2007). According to Abdi (2007), Type 1 error occurs due to an inflation of the alpha level. As a result the Sidak method is used to correct the alpha level making it more stringent, consequently resulting in less error (Abdi, 2007). From the resulting analysis, we can see (Table 16) that principals preferred option 3 (Implementing Lesson Plans) as a reason for teacher ineffectiveness relative to options 1 (Content Knowledge), 4 (Rapport), and 5 (Discipline), and they also preferred option 2 (Writing Lesson Plans) to option 1 (Content Knowledge). Other differences have confidence intervals that clearly span 0.

Table 16

Ordinal Probit Pairwise Differences

Dimension	vs	Dimension	Difference	Std. Err.	Lower CI	Upper CI
CM	vs	ILP	-0.61	0.15	-1.03	-0.18
Rapport	vs	ILP	-0.60	0.17	-1.07	-0.13
CM	vs	WLP	-0.40	0.18	-0.90	0.10
Rapport	vs	WLP	-0.38	0.18	-0.89	0.12
CM	vs	Rapport	-0.01	0.07	-0.21	0.18
CM	vs	CK	0.07	0.17	-0.41	0.54
Rapport	vs	CK	0.08	0.18	-0.42	0.58
ILP	vs	WLP	0.21	0.14	-0.19	0.61
WLP	vs	CK	0.46	0.17	-0.01	0.94
ILP	vs	CK	0.68	0.16	0.22	1.13

Note. Number of groups = 90. Number of Observations = 1,294

This order of perceived causes of teacher ineffectiveness differs from Torff and Sessions' (2009) findings when high school principals themselves were surveyed in the context of ELA and Math instruction. Once again, there were similar rankings in both studies for Content Knowledge, which was rated as the least frequent perceived cause of teacher ineffectiveness. When compared to Torff and Sessions' (2009) rankings, however, Implementation of Lesson Plans and Writing Lesson Plans replaced Classroom Management and Rapport with Students as the most frequently endorsed perceived causes of teacher ineffectiveness by the current, elementary school sample. Table 17 below demonstrates a comparison of the current study with Torff and Sessions (2009).

Table 17

Comparison of Perceived Causes of Teacher Ineffectiveness in the Current Study and in the Study by Torff and Sessions (2009) - Most Frequent to Less Frequent

Torff and Sessions (2009)	Current Study
Classroom Management	Implementing Lesson Plans
Rapport with Students	Writing Lesson Plans
Implementing Lesson Plans	Rapport with Students
Writing Lesson Plans	Classroom Management
Content Knowledge	Content Knowledge

Thus, HO3 is not supported, because the ordering of perceived elementary school principals' ratings of General causes of teacher ineffectiveness are different than the ordering of ratings obtained from Torff and Sessions' (2009) high school principal sample.

Summary

Table 18 summarizes the hypothesis testing results. Two of the four hypotheses received support.

Table 18

Overview of Results of Hypothesis Testing

Hypothesis	Study Hypothesis	Supported/ Not Supported
H01	Consistent with Torff and Sessions' (2005) findings, elementary school principals' general causes of teacher ineffectiveness will be (in order from the most frequent causes to least frequent causes): classroom management skills (pedagogical), lesson- implementation skills (pedagogical), rapport with students (pedagogical), lesson-planning skills (pedagogical), and content knowledge.	Not Supported
H02	Elementary school principals' perceived causes of teacher ineffectiveness within specific content areas (Math and ELA) will be different from each other.	Not Supported
H03	Consistent with Torff and Sessions' (2009) findings, elementary school principals' perceived causes of teacher ineffectiveness, within the context of ELA and Math instruction, will be (in order from the most frequent causes to least frequent causes): classroom management skills (pedagogical), rapport with students (pedagogical), lesson-implementation skills (pedagogical), lesson-planning skills (pedagogical), and content knowledge.	Not Supported

Qualitative Results

A total of 78 (75%) out of a possible 104 participants answered all or part of the qualitative survey questions (See Appendix I for detailed description of coding definitions for each question). The following sections provide descriptive statistics for each question, Cohen's kappa, as well as a brief summary of results. Percentages are calculated using an *n* of 78. See Appendix J for percentages of coded categories for each separate coder.

Question 14. What information goes into your evaluation of teachers (i.e., Do you use a rubric? Is it based on observation? Do you review lesson plans and student achievement data? How many times per year do you conduct an evaluation? Do you evaluate across different content areas?)?

Overall, coder A and B demonstrated strong coding agreement on Question 14 (Kappa = .92). Results indicate that over half of the participants who responded use a published evaluation system that includes a standards based rubric (Coder A= 56.41% and Coder B= 55.13%). Approximately another third of the participants utilizes some form of a rubric to evaluate teachers, however, the rubric is not a published system (Coder A= 30.77% and Coder B= 29.49%). Finally, according to both Coder A and B, only 12.82% of participants use their own teacher evaluation system that does not include any form of rubric and, in many cases, is entirely observation based. Table 19 includes frequencies and percentages for each response category.

Table 19

Frequencies and Percentages for Each Response Category for Question 14

Response Category	Frequency	Percentage
Option 1: The participant responds by indicating that he/she utilizes some form of an established standards based rubric. Examples include: Danielson, Marshall and/or NYSUT (New York State Union of Teachers) evaluation rubrics.	44	56.41
Option 2: The participant responds by indicating that he/she utilizes a rubric, however, does not reference any published or established system.	24	30.77
Option 3: The participant responds by indicating that they use some form of their own system (personal or school developed) that is not rubric based. This may include a participant who just indicates that they conduct observations, however, does not make reference to any pre-determined rating criteria.	10	12.82
Option 4: N/A	0	0

$n = 78$

Question 15. Do you believe your current evaluation system is a fair way to assess teachers? Why or why not?

Overall, coder A and coder B demonstrated strong agreement on Question 15 (Kappa = 0.87). Results indicate that approximately half of the participants who responded believe that their current evaluation system is a fair way to assess teachers (Coder A = 52.56% and Coder B = 50%). Out of the remaining participants, who responded "No", approximately one quarter indicated that they did not feel as if their current systems are fair because more information is needed, i.e. more observations, to make a valid evaluation (Option 4: Coder A = 25.64% and Coder B = 24.40%). The next most frequent reason for choosing "No" related to the use of student standardized test

information. About 10% of the answers to the question (Coder A = 10.26% and Coder B = 11.92%) indicated that more diverse student achievement information should be used and/or standardized test information is not a fair way to evaluate effective teaching practice. Table 20 includes frequencies and percentages for each response category.

Table 20

Frequencies and Percentages for Each Response Category for Question 15

Response Category	Frequency	Percentage
Option 1: Participant responds by indicating "Yes".	40	51.28
Option 2: Participants responds by indicating "No" AND identifies the reason as related to the use of student achievement data in the form of standardized testing. The participant may also indicate that more student achievement information is needed in various forms (i.e. individual student progress or social/ emotional learning).	9	11.54
Option 3: Participants responds by indicating "No" AND identifies the reason as related to not being able to effectively evaluate tenured teachers and/ or not allowing for tenure to be re-visited.	2	2.56
Option 4: Participants responds by indicating "No" AND identifies the reason as related to a validity concern due to more observations needed and/ or not enough information being provided (i.e. systems that only report a Satisfactory or Unsatisfactory rating).	20	25.64
Option 5: N/A	7	8.97

$n = 78$

Question 16: What other information that is not currently included in your evaluation process would you like in order to support your assessment of teachers?

Overall, coder A and coder B demonstrated strong agreement on Question 16 (Kappa = 0.87). Participants responded to Question 16 by citing various forms of information they would include in their current assessment. In fact, Options 1 – 4 ranged from approximately 11 -15% principal endorsement as evaluated across both coders. These options included professionalism, needing more support/time to conduct evaluations, roles of the teacher other than direct instruction as well as the inclusion of more diverse student achievement information. In addition it should be noted that approximately 40% of the participants who responded did not indicate they would include more information in addition to their current system (Coder A = 44.87% and Coder B = 41.03%). Table 21 includes frequencies and percentages for each response category.

Table 21

Frequencies and Percentages for Each Response Category for Question 16

Response Category	Frequency	Percentage
Option 1: The participant responds by indicating information related to professionalism. This may include being on time and/or dress code.	9	11.54
Option 2: The participant responds by indicating that he/ she needs more support or time to conduct evaluations. Support examples include further training in their evaluation systems, third party specialist involvement, and/ or support managing collected data.	10	12.82
Option 3: The participant responds by indicating information related to other roles of the teacher that is not under the umbrella of direct instruction. Examples may include lesson planning, discipline/ rapport building skills, involvement in school climate or extracurricular activities, etc.	10	12.82
Option 4: The participant responds by indicating information related to the inclusion of more diverse student achievement information. Examples include portfolios, longitudinal data, more frequent localized progress monitoring data, etc.	8	10.26
Option 5: The participant responds by indicating information related to feedback from parents and/ or a reference to a self-evaluation.	2	2.56
Option 6: N/A	33	42.31

Note. 6 participants (7.69%) left this question blank. $n = 72$

Question 17. What do you wish you could do in regard to teacher assessment? Hypothetically, what would you do if you were allowed to evaluate teachers in any way you believe is best?

Overall, coder A and coder B demonstrated strong agreement on Question 17 (Kappa = 0.91). Despite the fact that 40% of participants who responded to Question 16 indicated that they would not add any information to their existing systems, approximately 66% of participants responded to Question 17 indicating that they would like more information on which to base their evaluation of teachers. More specifically, approximately one third of the responses to Question 17 indicated that principals would like to increase the amount of informal teacher assessment and/or have more time allowed to spend in the classroom (Coder A and B = 32.05%). The next highest cited idea described by principals is be to utilize more diverse student achievement information (Coder A = 16.46% and Coder B = 19.23%), followed by principals' desire to be permitted to use teacher evaluation scores to more easily dismiss ineffective teachers (Coder A = 7.69% and Coder B = 6.41%). The remaining options (2, 3, and 6) accounted for 5% of less of principal responses. These included the ability to use technology, teacher content area assessment as well as teacher preparation assessment. It should be noted that approximately one quarter of the participants who responded did not have any further ideas about how to conduct teacher evaluation more effectively (Coder A = 29.49% AND Coder B = 25.64%). Table 22 includes frequencies and percentages for each response category.

Table 22

Frequencies and Percentages for Each Response Category for Question 17

Response Category	Frequency	Percentage
Option 1: The participant responds by indicating information related to the inclusion of more diverse student achievement information (may also indicate the removal of state wide standardized testing data). Examples of additions to an existing system may include student portfolios, longitudinal data, more frequent localized progress monitoring data, assessment of students in non-testing grades, social/ emotional learning assessment, etc.	15	19.23
Option 2: The participant responds by indicating information related to increasing the amount of observations allowed and/ or the time spent in the classroom directly working with/ coaching teachers.	25	32.05
Option 3: The participant responds by indicating information related to utilizing technology such as audio/ video recording to assist in the evaluation process.	2	2.56
Option 4: The participant responds by indicating information related to using evaluation results to dismiss ineffective teachers.	6	7.69
Option 5: The participant responds by indicating information related to teacher content area assessment.	2	2.56
Option 6: The participant responds by indicating information related to teacher preparation (i.e. Lesson Planning).	4	5.13
Option 5: N/A	22	28.21

Note. 2 participants (2.56%) left this question blank. $n = 76$

Summary of qualitative results. A review of the qualitative results reveals that the majority of principals sampled are currently using standards based rubric systems to evaluate teachers. Many of these systems are based on the work and research done by Danielson (2013). Even though these systems are comprehensive, about half of all principals do not believe that their system is a fair way to evaluate teachers. Most principals believed that they needed more information to make a valid assessment of teacher performance, and wanted to use less standardized test information as an indicator of teacher effects on student progress. Overwhelmingly, principals believed that they needed to conduct more observation, in particular more informal observation, to make better judgments of teacher performance. Many principals expressed the desire to have more time in the classroom, not only to conduct observation, but also to support teachers in their teaching practices. Consistent with previous responses, principals also expressed the need for more diverse student achievement information to be included in the evaluation progress. Principals demonstrated concern over the reliance on standardized test information and would like to see other forms of student assessment, such as portfolios and short-term progress monitoring measures included in the teacher evaluation process.

These results provide insight into the factors that principals believe are most important in evaluation. Instead of a standards based rubric, it appears as if many principals have confidence in their evaluation judgment if given the opportunity to conduct more observation. They have expressed a need for more exposure and time with a teacher. In addition, too much use of standardized testing information appears to be a common complaint shared by many principals. The implication is that progress made by

a student is not accurately captured by standardized assessments that in turn do not reflect the true impact of the teacher.

Chapter V

Discussion

The following section provides a brief summary of findings as well as how they fit with the existing literature. In addition, the chapter presents educational/school psychology implications, study limitations, and suggestions for future research.

Summary of Findings

The current study is an extension of Torff and Sessions' (2005, 2009) studies that investigated principals' perceived causes of teacher ineffectiveness using a sample of high school principals. The current study extended the research by giving Torff and Sessions' survey to a sample of New York State elementary school principals. Results indicate that, contrary to expectations, elementary school principals' rankings of the perceived causes of teacher ineffectiveness did not replicate the rankings by Torff and Sessions' high school principals. Principals in both samples, however, rated Content Knowledge as the least frequent perceived cause of teacher ineffectiveness. This study's elementary school principals rated Implementing Lesson Plans as the most frequent perceived cause of general teacher ineffectiveness followed by Writing Lesson Plans.

In addition, the current study extended the study by Torff and Sessions (2005, 2009) by obtaining elementary principals' ratings of the perceived causes of teacher ineffectiveness in the academic Domains of ELA and Math. Thus, the current study examined contextual factors (i.e., academic content area Domains). Results of model testing indicated that Dimension, or rating criterion, was the best predictor of principal perceptions. In other words, elementary principals' ratings varied only relative to the five perceived causes of teacher ineffectiveness. Academic domain did not significantly

relate to principals' ratings. This finding was consistent with results of Torff and Sessions (2009), because there was no significant Domain variability in either study. The actual rank order of perceived causes of teacher ineffectiveness, however, did vary between Torff and Sessions' (2009) sample of high school principals and the current sample of elementary school principals. Ordinal probit analysis, followed by general pairwise comparisons, determined that Implementation of Lesson Plans and Writing of Lesson Plans were the most frequent principal-rated causes of teacher ineffectiveness across all Domains (i.e., General, ELA, and Math).

Results in Context

As indicated above, results of the current study were similar to those obtained by Torff and Sessions (2005, 2009) in that Content Knowledge was rated as the least frequent cause of general teacher ineffectiveness. Thus, the results of all studies suggest that, in general, both elementary and high school principals believe that teachers have adequate knowledge of the subjects that they teach. The current study extended this finding to the specific content Domains of ELA and Math within an elementary school sample. Another similarity between the two studies is that principals in both samples reported Lesson Implementation Skills as one of the most frequent general causes of teacher ineffectiveness, ranked first by current elementary school principals and second by Torff and Sessions' (2005) high school principals. Torff and Sessions define Lesson Implementation Skills as effective execution of planned learning experiences during an instructional period. Hence, combined results of the two studies appear to indicate that principals believe that, even though teachers have Domain competence, they may be ineffective because they are unable to convey their knowledge to students in a

satisfactory manner.

Results indicate, however, that there were also differences between elementary and principals' perceptions of the causes of teacher ineffectiveness. Elementary school principals ranked Writing Lesson Plans as the second most frequent ineffectiveness cause. Coupled with their first place ranking of Implementing Lesson Plans, current results suggest that principals in this sample believed that ineffective elementary school teachers might need additional training in these two Dimensions.

Torff and Sessions' (2005) sample of high school principals rated Classroom Management Skills (ranked first) and Rapport with Students (ranked third), and Torff and Sessions' (2009) sample rated Classroom Management (ranked first) and Rapport with Students (ranked first) higher than did the current sample of elementary school principals, who ranked Discipline Skills third and Rapport with Students fourth. These different findings may point to an inherent difference between the skills necessary to be an effective high school and an effective elementary school teacher. A study by the National Comprehensive Center for Teacher Quality and Public Agenda (Rochkind, Ott, Immerwahr, Doble & Johnson, 2007) investigated differences between new teachers in high school and elementary schools. A sample of 641 first-year teachers from throughout the United States answered questions that covered issues relating to teacher training, recruitment, professional development, and retention. The results revealed differences between elementary and secondary school teachers' perceptions of their teaching experience and preparation that are consistent with differences in principals' rankings of the causes of teacher ineffectiveness between the current study and that of Torff and Sessions (2005, 2009).

Taken together, current results and those of Torff and Sessions (2005, 2009) suggest that practical classroom issues, such as classroom management, are a more important cause of teacher ineffectiveness for high school teachers than for elementary school teachers. Results of Rochkind et al.'s (2007) study suggest that teachers might agree with this statement. Rochkind et al. found that 53% of their sample of new secondary school teachers said their preparation was too theoretical versus 40% of new elementary school teachers. In addition, 88% of secondary school teachers surveyed indicated that the most pressing problems facing high schools are social problems as well as 'kids who misbehave'. Finally, according to Rochkind et al., 51% of high school teachers surveyed specified unmotivated students as a major drawback of teaching, versus 25% of elementary school teachers. Thus, Classroom Management Skills (as identified in Torff and Sessions (2005, 2009) appear to be a set of skills that high school teachers as well as high school principals in the current study identify as a deficit area. In contrast, Rochkind et al.'s elementary school teachers did not identify discipline skills as an area of need, particularly when asked about classroom management in reference to their teacher preparation programs. In support of their findings, the current elementary school principals, regardless of school demographic information, ranked classroom management skills third (out of five possible Dimensions) as a perceived cause of teacher ineffectiveness.

The current study results revealed that elementary school principals tended to perceive causes of ineffective teaching consistently across the possible confounding variables, namely instructional content areas ratings, that the study examined. Readers are cautioned, however, that this study did not examine all possible confounding

variables. Previous studies have indicated that student level variables such as socio-economic, minority, and/or non-English language status are significantly negatively related to student outcomes (Darling-Hammond, 1999; Espelage et al., 2013; Haertel, 1986). Others have shown that school-wide variables such as quality of materials, location of school, and school climate are also related to student achievement outcomes (Gottfredson, Gottfredson, Payne, & Gottfredson, 2005). A recent survey by Markow et al. (2013) revealed that principals in schools with two-thirds low-income students are less likely to give teachers excellent ratings compared to principals from schools with one-third or fewer lower income students.

It can be argued that teachers who work in schools with students who have profiles that correlate negatively with academic outcomes may have to adjust their teaching practices to address more pertinent needs. Recently Espelage et al. (2013) presented results of a nationwide poll that indicated that approximately 80% of teachers have reported at least one threat or violent incident toward them in the past year. In addition, of those teachers who reported an incident, 94% indicated that they were victimized by a student. This study is a poignant example of how teacher safety needs may need to be addressed to enable teachers to teach effectively. Espelage et al. indicated the need for teacher training in classroom management skills. In line with Espelage et al., Torff and Sessions' (2005, 2009) results showed that high school principals ranked Classroom Management Skills as the primary cause of teacher ineffectiveness; however, current elementary school principals did not view Classroom Management Skills with the same importance. Clearly, there should be further study of principals' perceived causes of teacher ineffectiveness that include large, diverse samples.

Results of previous research also indicate the importance of cross content domain ratings. Gallagher (2004) showed that teacher evaluation scores for literacy instruction explained 34% of variation in student test scores; however, teacher evaluation scores for math instruction were not a significant predictor of value added effects. Gallagher concluded this by conducting hierarchal linear modeling to estimate value-added teacher effects by correlating them with teacher evaluation scores in literacy, math, and language arts as well as a composite measure of student achievement. The teacher evaluation process used in Gallagher's study was an adaptation of the work by Danielson (1996) that includes a detailed rubric for principals to use to evaluate teachers that describes a level of proficiency for up to 9 standards within each of 10 domains (lesson planning, classroom management, literacy, mathematics, language development, special education inclusion, social studies, science, art, and technology) (Gallagher, 2004). These results suggest that effective math instruction may incorporate different teaching practices compared to literacy instruction; however, these practices were not identified successfully by principals even when using a detailed standards based rubric. Similarly, principals in the current study did not differ in their perceived causes of teacher ineffectiveness across academic content Domains.

As discussed, the results of the current study consistently indicated that the Dimensions of Implementing Lesson Plans and Writing Lesson Plans were the most frequently principal-rated cause of teacher ineffectiveness even when accounting for principal rating bias as well as potential differences between instructional content area ratings. This suggests that principals were possibly not considering other variables, such as individual students' needs, when conducting completing the survey. Previous research

(Haertel, 1986; Markow et al., 2013) suggests that this lack of consideration may not be the fault of the principal as a rater, but instead a result of the current emphasis on standardized test scores as well as successful implementation of the Common Core Standards.

The Common Core State Standards Initiative (2012) was developed by the National Governors Association (NGA) and the Council of Chief State School Officers (CSSO) and is considered part of a standards-based education reform that has been adapted by 45 states, including New York. The standards are considered evidenced based that, if mastered, lead to higher education and workforce readiness. Title 1 funds are contingent on full adoption of the Standards by 2015. The adoption of the Standards has already begun to take place as the New York State standardized tests in ELA and Math have been aligned to the Common Core Standards. In New York, all state assessments were aligned to the new Common Core Standards in the 2011-2012 school year and in the 2012-2013 school year, all grades 3-8 ELA and Math instruction needed to be aligned to the Common Core Standards (Changes to New York State Standards, 2013). These changes understandably may cause principals to emphasize teachers' abilities to Implement Lesson Plans as well as Write Lesson Plans effectively. Successful implementation of the Common Core Standards almost requires principals to focus only on these two Dimensions (i.e., Implementing Lesson Plans and Writing Lesson Plans) and their alignment to the Common Core. Readers should note that Common Core Standards were not in effect when Torff and Sessions (2005, 2009) conducted their study, and this may explain some differences in their results relative to those of the current study.

Previous research has explored how focus on meeting requirements of standardized

tests may impact teaching practice as well as teacher evaluation. Haladyna et al. (1991) identified practices such as developing a curriculum to match the test and preparing objectives to match the test as well as using commercial materials specifically designed to improve the test that may result from adoption of the Common Core Standards. Although these practices support schools' and teachers' alignment to the Common Core, they may also lead to unintended negative consequences. According to Haladyna et al. (1991), adoption of the Common Core Standards may limit the materials and methods by which teachers instruct their students, sacrificing teacher autonomy and individual student needs.

Meeting requirements of standardized testing is an issue in teacher evaluation identified by principals in the current study. Qualitative results indicated that 10%-11% of respondents do not believe their current system is fair due to the use of standardized testing information. In addition, 16%-19% of respondents, if given a choice, would utilize more diverse student achievement information and progress-monitoring data in replace of standardized testing information when evaluating teacher effectiveness.

Qualitative results also shed light on the importance of principal ratings and observation in the teacher evaluation process. Over 50% of the respondents indicated that they currently use an established/published standards based teacher evaluation system to conduct their teacher evaluations. As discussed earlier, principal observation is a major component in these systems. One example of such systems is Danielson's Framework (2013), which was adapted by the NYC DOE. According to McAdoo (2013), within this system, principal observation accounts for 60% of the teacher evaluation score (based on a 100 point system where 60 points is based on principal observation). Despite

the fact that a large portion of teacher evaluations are currently determined by principal observations, principals themselves indicated that they believe more observation, in particular more informal observation, is needed to make a more accurate and valid assessment of teacher performance. According to the qualitative results, 25% of the respondents indicated that they did not believe their current evaluation system is effective due to a need for more informal observations. In addition, 32% of respondents indicated that, if they had a choice, more informal observation and/or time spent within classrooms would be included in the teacher evaluation process.

Limitations

The most notable limitation of the current study is sample size. Torff and Sessions' (2005) survey of New York State high school principals included 242 participants and Torff and Sessions (2009) included 251 principals. The current study had 90 respondents who completed General ineffectiveness ratings, and only 83 principals who completed the entire survey. This most likely affected the robustness of the results. For example, mean differences in ratings were only found between the most frequently endorsed causes of perceived teacher ineffectiveness (Implementing Lesson Plans Effectively and Writing Lesson Plans). Although there were mean ratings differences among the other perceived causes of teacher ineffectiveness, the differences were not significant, but may have been with a larger sample.

Dimensional differences across content Domains may also have achieved significance with a larger sample. Principals' ratings of perceived causes of teacher ineffectiveness were different for Math instruction relative to their rankings for the General and ELA Domains. Content Knowledge, which was the least frequently

endorsed cause of teacher ineffectiveness in General and ELA instruction, was not the least frequently endorsed cause of Math teacher ineffectiveness. Despite this shift in ratings across Domains, subsequent analysis revealed that the variance explained by Domain was negligible and that the Domain variable provided little predictive value. A larger sample size may allow better detection of possible differences across Dimension and Domain ratings as demonstrated by the effect size reported for the difference in ratings between ELA and Math (g = .24). According to Cohen (1992), this effect size falls between small (.10) and medium (.30). According to Cohen (1992), a larger sample size may in fact demonstrate statistical difference for Content Knowledge in Math. A sample size of approximately 393 would yield significance at the $p < .05$ level for a small effect size. In addition, a sample size of 64 per group would be needed for significance for a medium effect size. One can then conclude that sample sizes similar to those of Torff and Sessions (2005, 2009) would be able to detect statistical differences in Content Knowledge as a perceived cause of teacher ineffectiveness in Math when compared to ELA.

Another limitation is the survey instrument itself. It only includes five possible Dimensions that reflect causes of perceived teacher ineffectiveness. As a result, principals are forced to rate only these five, excluding other, possibly more salient, causes. The five Dimensions may actually contain within them more specific perceived causes, which would enable a more detailed evaluation of teacher ineffectiveness. For example, the category most frequently endorsed as the leading perceived cause of teacher ineffectiveness, Implementing Lesson Plans, may include a plethora of teacher characteristics. One principal may believe that Implementing Lesson Plans primarily

means that a teacher is able to meet the stated objective of the planned lesson, while another principal may be looking for a teacher who effectively differentiates instruction. These differences may be dependent on the philosophy of the principal and/or the needs of the students. Regardless, they can both fall under the umbrella term of Implementing Lesson Plans; however, each may result in different observable teaching practices.

In addition, a third limitation of the current study is that student achievement information only included New York State-wide student achievement on Math and ELA tests. Mean scaled scores were used to obtain a general description of school level student achievement. Although these scores provided indicators of student achievement, there are some weaknesses associated with using state test data as the only indicator. The first is that schools in the current study varied in grade and subsequently, differed in terms of what version of each test was administered. In addition, there are some schools that have certain grades that achieved higher than other grades. This may indicate that teachers in certain grades within a school are more or less effective; however, the use of mean scaled scores could not pick up these differences. Another area where the study was not able to identify achievement effects was schools that have a large special education population. These schools tend to have students who score lower on standardized tests. In fact, in some cases these schools may not have any reported scores because students identified as functioning within a low cognitive range would take New York State Alternate Assessment and not the standardized ELA and Math tests. In schools with large special education populations, student achievement is measured in alternative ways and could not be captured by the scores used in the current study.

Implications for School Psychologists

The major implications of this study include informing both current teacher evaluation systems as well as current teacher training programs. In addition, this study's results may inform school psychologists in their consultations with administrators on a systems level basis and with teachers. One possible implication, for school psychologists, is within the role of consultant in regard to Curriculum Based Measurement (CBM).

Findings from the Curriculum Based Measurement (CBM) literature provide possible ways to increase teachers' effectiveness in planning and implementing lessons. According to Fuchs, Fuchs, Hamlett and Stecker (1991), CBM is a standardized measurement system for routinely indexing student achievement in the school curriculum. In addition, Fuchs et al. further define CBM as a methodology used to identify key curricular dimensions and use those dimensions to sample, administer and score tests, and organize assessment information to formulate instructional decisions (p. 620). The use of CBM leads to more targeted lesson planning as well as lesson implementation. In addition, recent research on CBM has investigated the effects of CBM plus diagnostic feedback. According to Caipizzi and Fuchs (2005), CBM with the addition of diagnostic feedback helps teachers to modify their initial instructions, to plan specialized adaptations, and to increase the number of interventions implemented. Overall, the use of CBM with instructional feedback to teachers has led to positive gains in student achievement. For example, Fuchs et al. (1991) randomly assigned 33 elementary school teachers to one of three treatment groups: a control group, a CBM group with expert system instructional consultation, and a CBM group with no expert

system instructional consultation. The study was conducted within the context of math instruction, and the Math Operations Test was the measure of achievement. The expert system is a computer program designed to use CBM information, inputted by the teacher, to create recommendations for instructional adjustment as well as how to implement those adjustments during instruction. Fuchs et al. found that neither the control nor the CBM alone groups led to significant increases in student achievement. However, the combination of CBM with the addition of expert systems instructional consultation led to greater student achievement. The authors also found that the CBM plus instructional consultation group made more changes to their instructional planning and implementation, which resulted in the gains in achievement. These include: teachers' employment of self-talk methods, alternative algorithms instead of using ones that had failed previously, and mixed problems to enhance retention and encourage student motivation. Overall, Fuchs et al. (1991) concluded that:

> ...more varied and perhaps more sound instructional revisions, associated with the CBM/ consultation group, were related to statistically significant student achievement differences of practical importance (in terms of effect size). One might characterize the instructional adjustments of these more successful teachers as focusing on both what to teach and how to restructure instruction. (p. 636).

School psychologists play a primary role in the evaluation of students referred for special education programs. In addition to training in the administration and interpretation of student assessment results, school psychologists are taught to consider other variables that may impact students' academic performance. This unique training

can be adapted to the evaluation of school level programs as well as to fair teacher evaluations. Too often school psychologists are only used for the single purpose of student psycho-educational evaluations. The National Association of School Psychologists' (NASP) Model for Comprehensive and Integrated School Psychology Services (2010) contains two major competencies of the school psychologist: Student Level Services and System Level Services. Systems Level Services is an umbrella term that may include school psychologists as participants in the development of fair evaluation systems. According to NASP (2010), school psychologists can contribute to system level services based on their knowledge of systems structure, organization, and theory. NASP (2010) further identifies one of the professional practices associated with systems level services as:

> ...the evaluation of the outcomes of the classroom, building, and system initiatives and the implementation of decision-making practices designed to meet general public accountability responsibilities. (p. 6)

The development and implementation of teacher evaluation systems may be considered as part of the evaluation of building and system initiatives. The researcher is not suggesting that school psychologists should evaluate teachers; however, they may contribute to the development of teacher and program evaluation and implementation through interpretation, collection, and analysis of system level data.

Many schools have more one than one administrator, such as an assistant principal. To make sure administrators conduct observations accurately and in a similar way (i.e., achieve good inter-rater reliability between administrators), the school psychologist can facilitate administrator training in teacher observation. In addition,

school psychologists may support schools in organizing data to make determinations concerning the relationship of evaluation scores to student outcomes and to assess any unintended consequences or validity threats to the evaluation current system.

Results of the current study indicate that principals are not affected by other variables, such as content area, when identifying causes of teacher ineffectiveness. This may indicate that principals are primarily only looking for a discrete number of teacher characteristics when conducting their observations. School psychologists can assist administrators to more effectively conduct these observations because they have extensive training in conducting objective student observations. The school psychologist also has unique information regarding student learning profiles and individual student needs. This is especially true for students who have qualified for special education services. The school psychologist may be able to assist the principal in understanding the student profiles that may alter how a teacher designs and delivers instruction allowing for more accurate evaluations of effectiveness because the principal will have a better understanding of what to look for and also understand why a teacher may deliver instruction in an alternative way.

Directions for Future Research

Future research is necessary to better understand how to conduct teacher evaluation in order to ensure it is both a fair system as well as one that translates into positive student outcomes. The current study identified Implementing Lesson Plans and Writing Lesson Plans as the two highest rated areas of perceived causes of teacher ineffectiveness. As discussed in the Limitations section, these are broad categories that could be studied in more detail. Future research can focus on further defining these

categories. For example, sub categories such as Aligning Lesson Plan to Standards or Implementing Lesson Based on Student Need are more specific characteristics that would be interesting to study. Inclusion of sub categories would allow researchers to understand better what principals are looking for as well as what teaching characteristics are considered the most salient. In addition, the refinement of all five Dimensions of the current questionnaire would possible allow for more differences to be exposed both within and between principals' ratings.

Another proposed study extension would include a more comprehensive focus on special education factors. The current study only included the Percentage of Students with an IEP as a Level 2 variable under the school level demographic information. Future research may reveal important factors that to consider, or not consider, when principals evaluate special education teachers. Although many special education students are still required to take grade level state tests, they work simultaneously on achieving individualized goals according to their Individualized Education Plans (IEPs). Future research should examine whether principals emphasize different teaching characteristics for a special education teacher relative to a general education teacher. In addition, future research should investigate how attainment of IEP goals factors in the final evaluation of the teacher compared to students' progress toward state-wide learning standards. It might be expected that other pedagogical teaching characteristics such as classroom management may play a much larger role in the evaluation of effective teaching within a special education class compared to effective teaching in a general education classroom.

A final suggestion for future research is for a more multifaceted use of student achievement data. The current study used mean ELA and Math scaled scores for each

entire school. These scores are from the previous year (2011) from when the survey was administered as well as being utilized as a variable to define student achievement on a school wide level. Future research should use both state-wide test scores as well as localized assessment consistent with the new teacher evaluation system proposed by Danielson (2000). The use of multiple achievement data provides a more accurate assessment of student functioning. Collection of classroom level student achievement data would also allow researchers to connect teacher evaluation, and/or perceived causes of teacher ineffectiveness, to individual teachers to allow for a more detailed understanding of the validity of teacher ratings. For example, a teacher may have a class did not perform well on state-wide testing at the end of the school year. If researchers can track students' progress over the course of the year; they may be able to categorize the teachers' effectiveness in promoting student growth over the academic year. This information could then be compared to how the principals evaluate individual teachers. This procedure may reveal how principals rate teachers, what principals deem important, and if principal ratings reflect the needs of individual students and classrooms.

Conclusion

The purpose of the current dissertation is to contribute to the existing literature on teacher evaluation. More specifically, the study utilized principals' perceptions to identify what principals, who often evaluate teachers, believe are the most frequent causes of teacher ineffectiveness. For this dissertation, the researcher extended a study by Torff and Sessions (2005, 2009) by including elementary school principal perceptions and investigating whether differences exist in elementary school principals' perceptions when asked to rate teacher ineffectiveness across specific academic content areas

(Domains). Utilizing an ordinal probit model, followed by likelihood ratio testing and general pairwise comparisons, the researcher determined that the only variable that significantly predicted principal perception was Dimension (i.e., rating criterion). In addition, the results revealed that, when the researcher controlled for principals' propensity to use the scale in different ways, Implementation Lesson Plans and Writing Lesson Plans were the most frequently rated causes of teacher ineffectiveness across all Domains. This research will hopefully contribute to the existing literature and help decision makers and researchers alike in developing and implementing fair and effective teacher evaluation systems.

Appendix A

Survey Instrument

5/21/12

Dear School Administrator,
The following information will be used for a research project that is aimed at determining principals' perceptions of teacher ineffectiveness. The information will be added to a growing body of research in order to expand our knowledge and competence in teacher training as well as effective methods to evaluate teachers. All information will be kept confidential.

I write to ask for ten minutes of your time for a research questionnaire. Just answer the opinion questions below and return this page in the enclosed envelope by December 1st, 2012. Your responses are totally confidential and very much appreciated. Please contact me if you have questions. Thank you! Sincerely,
Steve Franklin, MS, NCSP
PhD student at the CUNY Graduate Center – Educational Psychology Program
SFranklin@gc.cuny.edu
(516) 633-1120
Survey adapted from: Torff and Sessions (2005)

Opinion Survey

Personal Information

1. Age: _____

2. Gender: (circle one) Male Female

3. Years as a classroom teacher: _____

4. Years as an administrator: _____

5. Educational attainment: (circle one) Bachelor's Master's

Master's + 30 Doctorate

School- Level Information

6. Grade levels in your school (i.e. K-5, K-8, etc.): _____

7. Location of school: (circle one) Rural Suburban Urban

8. Region in New York State (i.e. Hudson Valley, Southern Tier, etc.):
_____

8. Average class size: _____

9. Percentage of students who receive free or reduced lunch: _____

10. Percentage of students with an IEP: _____

General Perceived Causes of Teacher Ineffectiveness Ratings

12. When a teacher's classroom work is ineffective, it is because the teacher: (circle one for each statement)

fails to demonstrate needed content knowledge (teacher does not exhibit suitable expertise in the subject being taught)	**1** very rarely	**2** seldom	**3** sometimes	**4** frequently
fails to write effective lesson plans (teacher does not prepare appropriate learning experiences prior to an instructional period)	**1** very rarely	**2** seldom	**3** sometimes	**4** frequently
fails to implement lesson plans skillfully (teacher does not execute planned learning experiences effectively during an instructional period)	**1** very rarely	**2** seldom	**3** sometimes	**4** frequently
fails to establish sufficient rapport with students (teacher does not demonstrate adequate human relations and communication skills)	**1** very rarely	**2** seldom	**3** sometimes	**4** frequently
fails to maintain satisfactory classroom discipline (teacher does not successfully keep students on task and attentive)	**1** very rarely	**2** seldom	**3** sometimes	**4** frequently

13. When a teacher's classroom work is **ineffective**, it is because the teacher: *(check one for each subject)*

	Subject	*Very Rarely*	*Seldom*	*Sometimes*	*Frequently*
fails to demonstrate needed **content knowledge** (teacher does not exhibit suitable expertise in the subject being taught)	English	☐	☐	☐	☐
	Math	☐	☐	☐	☐
fails to **write lesson plans** that are effective (teacher does not prepare appropriate learning experiences prior to an instructional period)	English	☐	☐	☐	☐
	Math	☐	☐	☐	☐
fails to **implement lesson plans** skillfully (teacher does not execute planned learning experiences effectively during an instructional period)	English	☐	☐	☐	☐
	Math	☐	☐	☐	☐
fails to establish sufficient **rapport with students**(teacher does not demonstrate adequate human relations and communication skills)	English	☐	☐	☐	☐
	Math	☐	☐	☐	☐
fails to maintain satisfactory **classroom management** (teacher does not successfully keep students on task and attentive)	English	☐	☐	☐	☐
	Math	☐	☐	☐	☐

Open Ended Questions

14. What information goes into your evaluation of teachers? (i.e., Do you use a rubric? Is it based on observation? Do you review lesson plans and student achievement data? How many times per year do you conduct an evaluation? Do you evaluate across different content areas?)

15. Do you believe your current evaluation system is a fair way to assess teachers? Why or why not?

16. What other information, that is not currently included in your evaluation process would you like in order to support you assessment of teachers?

17. What do you wish you could do in regard to teacher assessment? Hypothetically, what would you do if you were allowed to evaluate teachers in any way you believe is best?

Appendix B

Sample Survey Monkey E-mail Sent to Principals

Message Preview
Below is a preview of your message based on the first recipient in your list ([Email]).

To: [Email]
From: "sfranklin@gc.cuny.edu via surveymonkey.com" <member@surveymonkey.com>

Subject: Survey regarding the rating of teacher effectiveness
Body: We are conducting a survey, and your response would be appreciated.

Here is a link to the survey:
https://www.surveymonkey.com/s.aspx

This link is uniquely tied to this survey and your email address. Please do not forward this message.

Thanks for your participation!

Appendix C

Consent Form

Dear School Administrator,

My name is Steven Franklin and I am a student in the Educational Psychology Ph.D. Program at The Graduate Center of the City University of New York (CUNY), and Principal Investigator of this project, entitled "Elementary School Principals' Perceptions of the Causes of Teacher Ineffectiveness and the Relation to Student Achievement Outcomes." This is a research study of principals' perceptions on the causes of teacher ineffectiveness and how/ if that relates to certain student achievement outcomes. The study is expected to gain an understanding of how to more effectively evaluate teachers as well as provide more guidance for teacher training programs. I would like you to fill out the survey. The survey includes directions and asks you to rate teacher ineffectiveness on 5 dimensions. This is then expanded to teacher competency across subject areas. Student achievement data will only include school performance on state-wide assessments. No individual student data will be used. Only information that is publicly available will be included and school identifying information will also be kept confidential.

This survey should not take any longer than 15-30 minutes. All information gathered will be kept strictly confidential, and will be electronically on my personal computer and be password protected. At any time you can refuse to answer any questions or not retract your survey after completing it.

The risks from participating in this study are no more than encountered in everyday life. The benefits of your participation are that a better understanding will be gained on more effective and fair ways to evaluate teachers. Furthermore, teacher-training programs will benefit from tailoring their instruction around the information gained from your participation. There will be approximately 300 of participants taking part in this study.

I may publish results of the study, but names of people, or any identifying characteristics, will not be used in any of the publications. If you would like a copy of the study, please provide me with your address and I will send you a copy in the future.

If you have any questions about this research, you can contact me at 516-633-112-or SFranklin@gc.cuny.edu, or my advisor Dr. Tryon at gtryon@gc.cuny.edu. If you have questions about your rights as a participant in this study, you can contact Kay Powell, IRB Administrator, The Graduate Center/City University of New York, (212) 817-7525, kpowell@gc.cuny.edu.

Thank you for your participation in the study. I will give you a copy of this form to take with you.

I agree to participate in the study by completing the survey: Yes No

_____ _____

Participant's signature Date

Appendix D

E-mail Correspondence Granting Permission to Use Survey from Torff and Sessions (2005)

To: Internal Review Board
From: Bruce Torff
Date: 9 June 2012

I hereby authorize Stephen Franklin to use the TAP survey in his dissertation research.

Bruce Torff

Bruce Torff, Ed. D.
Professor, Department of Teaching, Literacy, and Leadership
Director, Doctoral Program in Learning and Teaching

School of Education, Health and Human Services
Hofstra University
Hempstead, NY 11549
Phone: (516) 463-5803
Fax: (516) 463-6196
Email: Bruce. Torff@Hofstra.edu

Appendix E

ANOVA Tables for Principal Dimension Ratings Within Domains (General, ELA, and Math)

General Ratings	df	Sum of Squares	Mean Square	F
Dimension	4	16.68	4.17	6.63***
Residuals	444	279.31	0.63	

**p<0.001

ELA Ratings	*df*	Sum of Squares	Mean Square	*F*
Dimension	4	10.17	2.54	3.79**
Residuals	410	274.77	0.67	

**p < 0.01

Math Ratings	*df*	Sum of Squares	Mean Square	*F*
Dimension	4	6.90	1.73	2.72*
Residuals	410	259.77	0.63	

*p < 0.05

Appendix F

General Rating Pairwise Comparisons

General Rating Pairwise Comparisons- (Tukey HSD: 99% Confidence Interval)

Dimension	vs	Dimension	Difference	Std. Err.	Lower CI	Upper CI	Remarks
CK	vs	WLP	-0.38	0.11	-0.66	-0.10	WLP sig. higher
CK	vs	ILP	-0.57	0.11	-0.84	-0.29	ILP sig. higher
CK	vs	Discipline	-0.23	0.12	-0.53	0.08	NS
CK	vs	Rapport	-0.17	0.12	-0.48	0.15	NS
WLP	vs	ILP	-0.19	0.10	-0.44	0.06	NS
WLP	vs	Discipline	0.16	0.10	-0.11	0.43	NS
WLP	vs	Rapport	0.21	0.11	-0.08	0.50	NS
ILP	vs	Discipline	0.35	0.11	0.07	0.63	ILP sig. higher
ILP	vs	Rapport	0.40	0.11	0.10	0.67	ILP sig. higher
Discipline	vs	Rapport	0.06	0.16	-0.15	0.26	NS

Note. N= 90.

Appendix G

ANOVA Table between ELA and Math Ratings

ELA vs. Math Comparison	*df*	Sum of Squares	Mean Square	*F*
Domain	1	0.00	0.03	0.03 (NS)
Residuals	1262	842.90	0.67	

NS = Not Significant

Appendix H

Mixed Ordinal Probit Model – Fit Using Stata/ IC 13 Output

```
. meglm y i.domain##i.dimension || id:, family(ordinal) link(probit) vce(robust)

Mixed-effects GLM                       Number of obs     =       1294
Family:              ordinal            Number of groups  =         91
Link:                probit
Group variable:      id

Log pseudolikelihood = -1332.4617
Wald chi2(14)        =      45.13
Prob > chi2          =     0.0000
```

(Std. Err. adjusted for clustering on id)

y	Coeff.	Robust SE	z	P>z	95% CI
Domain					
ELA	.0329247	.1128218	0.29	0.770	-.1882019 - .2540513
Math	.3912103	.1320654	2.96	0.003	.13236680 - .6500538
Dimension					
WLP	.640428	.1809576	3.54	0.000	.28575750 - .9950984
ILP	.9427819	.1841652	5.12	0.000	.58182470 - 1.303739
Rapp.	.3345129	.1915434	1.75	0.081	-.0409052 - .7099309
CM	.2299211	.1958133	1.17	0.240	-.1538659 - .6137081
DomainXDime.					
ELAXWLP	-.0576302	.1336690	-0.43	0.666	-.3196167 - .2043563
ELAXILP	-.2534863	.1607919	-1.58	0.115	-.5686327 - .0616601
ELAXRapp.	-.2038014	.1601670	-1.27	0.203	-.5177229 - .1101202
ELAXCM	-.0456868	.1921306	-0.24	0.812	-.4222557 - .3308822
MathXWLP	-.4700195	.1472620	-3.19	0.001	-.7586477 - - .1813912
MathXILP	-.5450590	.1677097	-3.25	0.001	-.8737639 - - .2163540
MathXRapp.	-.5593723	.1868434	-2.99	0.003	-.9255787 - - .1931659
MathXCM	-.4414633	.1983938	-2.23	0.026	-.8303080 - - .0526186
ID					
var(_cons)	.5476173	.1307873			.3429138 - .8745193

Appendix I

Definitions for Coding Participants Responses for Qualitative Survey Questions

Question 14: What information goes into your evaluation of teachers? (i.e., Do you use a rubric? Is it based on observation? Do you review lesson plans and student achievement data? How many times per year do you conduct an evaluation? Do you evaluate across different content areas?)

Option 1: The participant responds by indicating that he/she utilizes some form of an established standards based rubric. Examples include: Danielson, Marshall and/or NYSUT (New York State Union of Teachers) evaluation rubrics.

Option 2: The participant responds by indicating that he/ she utilizes a rubric, however, does not reference any published or established system

Option 3: The participant responds by indicating that they use some form of their own system (personal or school developed) that is not rubric based. This may include a participant who just indicates that they conduct observations, however, does not make reference to any pre-determined rating criteria.

Option 4: N/A

Question 15: Do you believe your current evaluation system is a fair way to assess teachers? Why or why not?

Option 1: Participant responds by indicating "Yes".

Option 2: Participants responds by indicating "No" AND identifies the reason as related to the use of student achievement data in the form of standardized testing. The participant may also indicate that more student achievement information is needed in various forms (i.e. individual student progress or social/ emotional learning).

Option 3: Participants responds by indicating "No" AND identifies the reason as related to not being able to effectively evaluate tenured teachers and/ or not allowing for tenure to be re-visited.

Option 4: Participants responds by indicating "No" AND identifies the reason as related to a validity concern due to more observations needed and/ or not enough information being provided (i.e. systems that only report a Satisfactory or Unsatisfactory rating).

Option 5: N/A

Question 16: What other information, that is not currently included in your evaluation process would you like in order to support your assessment of teachers?

Option 1: The participant responds by indicating information related to professionalism. This may include being on time and/or dress code.

Option 2: The participant responds by indicating that he/ she needs more support or time to conduct evaluations. Support examples include further training in their evaluation systems, third party specialist involvement, and/ or support managing collected data.

Option 3: The participant responds by indicating information related to other roles of the teacher that is not under the umbrella of direct instruction. Examples may include lesson planning, discipline/ rapport building skills, involvement in school climate or extracurricular activities, etc.

Option 4: The participant responds by indicating information related to the inclusion of more diverse student achievement information. Examples include portfolios, longitudinal data, more frequent localized progress monitoring data, etc.

Option 5: The participant responds by indicating information related to feedback from parents and/ or a reference to a self-evaluation.

Option 6: N/A

Question 17: What do you wish you could do in regard to teacher assessment? Hypothetically, what would you do if you were allowed to evaluate teachers in any way you believe is best?

Option 1: The participant responds by indicating information related to the inclusion of more diverse student achievement information (may also indicate the removal of state wide standardized testing data). Examples of additions to an existing system may include student portfolios, longitudinal data, more frequent localized progress monitoring data, assessment of students in non-testing grades, social/ emotional learning assessment, etc.

Option 2: The participant responds by indicating information related to increasing the amount of observations allowed and/ or the time spent in the classroom directly working with/ coaching teachers.

Option 3: The participant responds by indicating information related to utilizing technology such as audio/ video recording to assist in the evaluation process.

Option 4: The participant responds by indicating information related to using evaluation results to dismiss ineffective teachers.

Option 5: The participant responds by indicating information related to teacher content area assessment.

Option 6: The participant responds by indicating information related to teacher preparation (i.e. Lesson Planning).

Option 5: N/A

Appendix J

Qualitative Respondent Percentages per Option for Each Coder

Question	Coder A	Coder B
14		
Option 1	56.41	55.13
Option 2	30.77	29.47
Option 3	12.82	12.82
15		
Option 1	52.56	50
Option 2	10.26	11.92
Option 3	2.56	2.44
Option 4	25.64	24.40
Option 5	25.64	7.70
16		
Option 1	15.38	11.54
Option 2	12.82	12.82
Option 3	11.54	12.82
Option 4	7.70	11.54
Option 5	2.56	2.56
Option 6	44.87	41.03
17		
Option 1	16.46	19.23
Option 2	32.05	32.05
Option 3	2.56	2.56
Option 4	7.69	6.41

Question	Coder A	Coder B
Option 5	3.85	2.56
Option 6	5.13	2.56
Option 7	29.49	25.64

Note. Percentages calculated by using the total number of respondents who responded to the Qualitative questions. N= 78.

References

Archibald, S. (2006). Narrowing on educational resources that do affect student achievement. *Peabody Journal of Education, 81,* 23-42. doi: 10.1207/S15327930pje8104_2

Achilles, C.M., & Finn, J.D. (2007). *Class-size policy: The star experiment and other class-size studies.* Buffalo, NY: University of Buffalo.

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R. Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard, L.A. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing paper 278). Washington, D.C.: Economic Policy Institute.

Beland, S., Klugkist, I., Raiche, G., & Magis, D. (2012) A short introduction into Bayesianevaluation of informative hypotheses as an alternative to exploratory comparisons of multiple group means. *Tutorials in Quantitative Methods for Psychology, 8,* 122-126.

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*, 3-20. doi: 10.1086/496904

Caipizzi, A.M. & Fuchs, L.S. (2005). *Effects of curriculum-based measurement with and Without diagnostic feedback on teacher planning.* Remedial and Special Education, *26,* 159- 174. doi: 10.1177/07419325050260030401

Campbell, D.T. (1976). *Assessing the impact of planned social change.* Hanover, NH: The Public Affairs Center, Dartmouth College.

Changes to New York State Standards, Curricula, and Assessments: ELA and Mathematics. (2013). Retrieved November 8, 2013, from http://www.engageny.org/sites/default/files/resource/attachments/ ccsstimeline.pdf

Chetty, R., Friedman, J., & Rockoff, J. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood.* Working Paper No. 17699 : Cambridge, MA: National Bureau of Economic Research.

Cohen, J. (1992). *A power primer.* Psychological Bulletin, 112, 155-159. doi: 10.1037/00033-2902.112.1.155

Danielson, C. (2013). The Framework for Teaching: Evaluation instrument. Princeton, NJ: The Danielson Group. Retrieved July 8, 2013, from http://www.danielsongroup.org/userfiles/files/downloads/2013EvaluationInstrument.pdf

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

Darling-Hammond, L. (1999). *Teacher quality and student achievement: A review of state policy evidence.* St. Louis, MO: Center for the Study of Teaching and Policy: University of Washington.

Darling-Hammond, L., Wise, A.E., & Klein, S. (1999). *A license to teach: Raising standards for teaching.* San Francisco, CA: Josey-Bass.

Darling-Hammond. L., & Snyder, J. (2000). Authentic assessment in teaching in context. *Teaching and Teacher Evaluation, 16,* 523-545. doi: 10.106/S0742-051X(00)

Descriptions of performance levels for the grades 3-12 English Language Arts Test. (2013). Retrieved April 13, 2013, from http://www.p12.nysed.gov/irs/ela-math/2012/2012-ELADefinitionsofPerformanceLevels.pdf.

Descriptions of performance levels for the grades 3-12 Mathematics Test. (2013). Retrieved April 13, 2013, from http://nysed.gov./irs/ela-math MATHDefinitionsofPerformanceLevels.pdf./2012/2012-2013

Dunn, T., Burman, J., & Briggs, J. (2010). Grade 3-8 Math and English Test results released: Cut scores set to new college-ready proficiency standards. *New York State Education Department.* Retrieved 3/31/12 from, http://www.oms.nysed.gov/press/Grade3-8_Results07282010.html

English Language Arts and Mathematics Assessment Results. (2013). Retrieved August 5, 2013, from http://www.p12.nysed.gov/irs/ela-math/.

Espelage, D., Anderman, E.M., Brown, V.E., Jones, A., Lane, K.L., McMahon, S.D....Reynolds, C.R. (2013). Understanding and preventing violence against teachers: Recommendations for a national research, practice, and policy agenda. *American Psychologist, 68,* 75-87. doi: 10.1037/a0031307

Farberman, R.K. (Ed.). (2013, November). Preventing Violence Against Teachers. *Monitor on Psychology, 44,* 58-66.

Fertig, Beth. (2012). Bloomberg 'optimistic' on teacher evaluations. *School Book.* Retrieved 3/31/12 from, http://www.nytimes.com/schoolbook/2012 /01/18/bloomberg-optimistic-on-teacher-evaluations/

Fuchs, S.L., Fuchs, D., Hamlet, C.L., S & Stecker, P.M. (1991). *Effects of curriculum based measurement and consultation on teacher planning and student achievement in mathematics operations.* American Education Research Journal, *28,* 617-641. doi: 10.3102/00028312028003617

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, 4,* 79-107. doi: 10.1207/S15372930pje7904_5

Glass, G.V. (1990). Using student test scores to evaluate teachers. In Jason Milman and Linda Darling-Hammond (Eds.), *The new handbook for teacher evaluation: Assessing elementary and secondary school teachers* (pp. 229-240). Newbury Park, CA: SAGE Publications.

Glass, G.V., & Hopkins, K.D. *Statisitcal methods in education and psychology* (3rd ed.). Boston, MA: Allyn and Bacon.

Gottfredson, G.D., Gottfredson, D.S., Payne, A.A. & Gottfredson, N.C. (2005). School climate predictors of school disorder: Results from a national study of delinquency prevention in schools. *Journal of Research in Crime and Delinquency, 42,* 412-444.

Gregory, R.J. (2004). *Psychological testing: History. principles, and applications* (4th ed.). Pearson: Boston.

Gross, M. (1999). *The conspiracy of ignorance: The failure of American public schools.* New York, NY: HarperCollins.

Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. *Educational Evaluation and Policy Analysis, 8,* 45-60. doi: 10.3102/01623737008001045

Hanushek, E.A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19,* 141-164. doi: 10.3102/01623737019002141

Hanushek, E.A., & Rivkin, S.G. (2010). *Using value-added measures of teacher quality.* Working Paper 9: Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, Urban Institute.

Haladyna, T.M. (1991). Raising standardized-achievement test scores and the origins of test score pollution. *Educational Researcher, 20,* 2-7. doi: 10.3102/0013189x020005002

Harris, D.N., & Sass, T.R. (2009). *What makes for a good teacher and who can tell?* Working Paper 30: Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, Urban Institute.

Heneman, H.G., Milanowski, A., Kimball, S.M., & Odden, A. (2006).*Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay.* (RB-45). Philadelphia, PA: Consortium for Policy Research in Education.

Individuals With Disabilities Education Improvement Act of 2004 (IDEA). Pub. L. 108-446, 118 Stat. 2647 (2004).

Kane, Thomas, J., Taylor, E.S., Tyler, J.H., & Wooten, A.L. (2010). *Identifying effective classroom practices using student achievement data.* Working Paper 15803: Cambridge, MA: National Bureau of Economic Research.

Koch, G. (1982). "Intraclass correlation coefficient". In Samuel Kotz and Norman L. Johnson. *Encyclopedia of Statistical Science.* New York: John Wiley and Sons, 213-217/

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis, 25,* 287-298. doi: 10.3102/01623737025003287

Larry P. v. Riles, No. C-71-2270-RFP (N.D. Cal., October 16, 1979).

Majority Press. (March 1, 2011). Senate passes bill to reform "Last in, First Out". *Majority Press.* Retrieved 4/21/11, from http://www.nysenate.gov/press-release/senate-passes-bill-reform-last-first-out

Markow, D., Macia, L., & Lee, H. (2013). *The MetLife survey of the American teacher: Challenges for school leadership.* New York, NY: MetLife. Retrieved from https//:www.metlife.com/assets/cao/foundation/MetLife – Teacher – Survey – 2012.pdf

McAdoo, M. (2013).*Teacher evaluation: Complex system unveiled.* UFT.org. Retrieved From http://www/uft.org/news-stories/complex-new-system-unveiled.

McAdoo, M. (2013). *Observations, with feedback, at heart of new plan.* UFT.org. Retrieved From http://www/uft.org/news-stories/assessing-student-progress.

Medley, D.M., & Coker, H, (1987). The accuracy of principals' judgments of teacher performance. *The Journal of Educational Research, 80, 242-247.*

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

NASP. (2010). *Model for Comprehensive and Integrated School Psychological Services.* Retrieved 11/8/2013, from http://www.nasponline.org /standards/2010standards/2_PracticeModel.pdf

New Jersey Department of Education. (September 1, 2011). *Department of Education announces 11 districts to participate in a teacher evaluation pilot program.* Retrieved 2/20/12. from http://www.state.nj.us/education/-news/2011/0901ee4nj.htm

No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2003).

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

Peterson, K. (2004). Research on school evaluation. *NASSP Bulletin, 88,* 60-79. doi: 10.11770/019263650408863906

Popham, W.J. (1999). Why standardized tests don't measure educational quality. *Using Standards and Assessment, 56,* 8-15.

Ravitch, D. (2000). *Left back: A century of failed school reforms.* New York, NY: Simon & Schuster.

Rochkind, J., Ott, A., Immerwhar, J., Doble, J. & Johnson, J. (2007). *Lessons learned: new teachers talk about their jobs, challenges and long range plans: A report from the National Comprehensive Center for Teacher Quality and Public Agenda.* New York: Public Agenda.

Rothstein, J. (2009). *Teacher quality in educational production: Tracking, decay and student achievement.* Unpublished manuscript.

Rothstein, L.F. (1995). *Special education law (3rd ed.).* New York: Longman.

Skiba, R., Knesting, K., & Bush, L.D. (2002). Culturally competent assessment: More than nonbiased tests. *Journal of Child and Family Studies, 11,* 61-78. doi: 10.1023/A:1014767511894

Santos, F. (2011). Dispute over evaluations imperils grants for schools. *The New York Times.* Retrieved from, http://www.nytimes.com/2011/12/31/-education/teacher-evaluations-dispute-imperils-grants-for-schools.html?_r=1

Sartain, L., Stoelinga, S.R., & Brown, E.R. (2011). Rethinking teacher evaluation in conferences and district implementation. *Consortium on Chicago School Research,* 1-62.

Sattler, J.M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego: Jerome M. Sattler, Publisher, Inc.

Schochet, P.Z., & Chiang, H.S. (2010). *Error rates in measuring teacher and school performance based on student test score gains, (NCEE 2010-4004).* Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Stodolsky, S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher, 13,* 11-18.

Stodolsky, S. (1990). Classroom observation. In Jason Milman and Linda Darling-Hammond (Eds.), *The new handbook for teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175-190). Newbury Park, CA: SAGE Publications.

Taylor, E.S., &Tyler, J.H. (2011). *The effect of evaluation on performance: Evidence from longitudnal student achievement data of mid-career teachers.* National Bureau of Economic Research Working Paper 16877.

Torff, B., & Sessions, D.N. (2005). Principals' perceptions of the causes of teacher ineffectiveness. *Journal of Educational Psychology,* 530-537. doi: 10.1037/0022-0663.97.4.530

Torff, B., & Sessions, D. (2009). Principals' perceptions of the causes and teacher ineffectiveness in different secondary subject. *Teacher Education Quarterly, 36*(3), 127-148. doi: 10.1080/00131725.2010.528553

U.S. Department of Education. (2009). *Race to the Top Fund – Executive Summary: Notice of proposed priorities, definitions, and selection criteria.* Washington, D.C.: U.S. Department of Education.

Wilkerson, D.J., Manatt, R.P., Rogers, M.A., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360° Feedback® for teacher evaluation. *Journal of Personnel Evaluation in Education, 14,* 179-192. doi: 10.1023/A:1008158904681

Wright, S.P., & Sanders, W.L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personal Evaluation in Education, 11,* 57-67. doi: 101023/A:1007999204543